IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 4, 2021, accepted January 12, 2021, date of publication February 12, 2021, date of current version March 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059343

Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review

FARRUKH ASLAM KHAN[1], (Senior Member, IEEE), KHAN ZEB[2], MABROOK AL-RAKHAMI[3,4], (Member, IEEE), ABDELOUAHID DERHAB[1], AND SYED AHMAD CHAN BUKHARI[5], (Senior Member, IEEE)

[1]Center of Excellence in Information Assurance, King Saud University, Riyadh 11653, Saudi Arabia
[2]Department of Electrical & Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada
[3]Research Chair of Pervasive and Mobile Computing, King Saud University, Riyadh 11653, Saudi Arabia
[4]Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11653, Saudi Arabia
[5]Division of Computer Science, Mathematics and Science (Healthcare Informatics), Collins College of Professional Studies, St. John's University, New York City, NY 11439, USA

Corresponding author: Farrukh Aslam Khan (fakhan@ksu.edu.sa)

This work was supported by the Deputyship for Research and Innovation, "Ministry of Education" in Saudi Arabia under Project IFKSURP-255.

ABSTRACT Diabetes is one of the most rapidly growing chronic diseases, which has affected millions of people around the globe. Its diagnosis, prediction, proper cure, and management are crucial. Data mining based forecasting techniques for data analysis of diabetes can help in the early detection and prediction of the disease and the related critical events such as hypo/hyperglycemia. Numerous techniques have been developed in this domain for diabetes detection, prediction, and classification. In this paper, we present a comprehensive review of the state-of-the-art in the area of diabetes diagnosis and prediction using data mining. The aim of this paper is twofold; firstly, we explore and investigate the data mining based diagnosis and prediction solutions in the field of glycemic control for diabetes. Secondly, in the light of this investigation, we provide a comprehensive classification and comparison of the techniques that have been frequently used for diagnosis and prediction of diabetes based on important key metrics. Moreover, we highlight the challenges and future research directions in this area that can be considered in order to develop optimized solutions for diabetes detection and prediction.

INDEX TERMS Diabetes, data mining, big data, prediction, detection, e-Health, m-Health.

I. INTRODUCTION

Diabetes is a chronic and non-communicable disease that destabilizes the normal control of blood glucose concentration in the body. The blood glucose concentration is usually regulated by two hormones, namely insulin and glucagon, which are secreted by beta (β) and alpha (α) cells of pancreas respectively [1], [2]. The normal secretion of both hormones sustains normal blood glucose concentrations in the body, which are in the range of 70 – 180 mg/dl (4.0 – 7.8mmol/L). Insulin decreases the level of glucose concentration, whereas glucagon increases it. However, abnormal secretion of these hormones leads to diabetes. There are a number of different types of diabetes with different prevalence; however, the most common types are type 1 diabetes, type 2 diabetes, and gestational diabetes mellitus (GDM). Type 1 diabetes commonly develops in children; type 2 diabetes is more prevalent in the middle-aged and elderly persons, while GDM appears in women and is diagnosed during pregnancy. In type 1 diabetes, the secretion of insulin fails due to the destruction of pancreatic beta cells, whereas in type 2, failures occur in both insulin secretion and action. GDM is a condition of glucose intolerance of any degree that is first recognized during pregnancy; mainly, it occurs in the second half of pregnancy. It can be mild, but it can also be associated with considerable hyperglycemia and high insulin requirements during pregnancy. All of these types result in unbalanced blood glucose concentration in the human body, which leads to severe health conditions in the body. Consequently, when the blood glucose concentration increases and exceeds the normal concentration range, then this condition is known as hyperglycemia. On the other hand, when it decreases and

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

becomes lower than the normal range, then such a condition is known as hypoglycemia [3]–[5]. Both of these conditions can lead to adverse consequences on an individual's health, for instance, hyperglycemia has long-term complications and can cause nephropathy, retinopathy, cardiovascular and heart diseases, and other tissue injuries, whereas hypoglycemia has short-term effects that may result in life-threatening diabetic coma [1], [3], [4].

Diabetes has become one of the major public health problems in today's world due to its prevalence in children as well as in the adult population. According to [6], [7], approximately 8.8% of the adult population was diabetic worldwide during 2015, which counts for around 415 million people, and is expected to reach around 642 million by 2040. In addition, the disease has affected more than half a million children during this period and has caused about 5 million deaths. On the other hand, in 2015, the estimated global economic burden of diabetes was nearly USD 673 billion, which is projected to be around USD 802 billion in 2040 [7].

Self-monitoring of blood glucose (SMBG) using finger-stick blood samples is a common approach of diabetes therapy that has been introduced three decades ago [8], [9]. In this approach, diabetics measure their blood glucose levels three to four times a day in an invasive way by pricking the skin of their finger using finger-stick glucose meters. The notion here is to collect blood glucose concentration levels at different times, and accordingly, adjust the insulin intake, diet, and exercise in order to maintain normal glucose levels. Nevertheless, this method is not only troublesome and painful but can also be misleading if the approximation of inulin intake is made based on merely few SMBG samples. Consequently, this could result in plasma glycemic levels to exceed the normal range. To overcome this problem, continuous glucose monitoring (CGM) has been introduced that provides maximal information about variations in blood glucose concentration throughout the day and enables optimal therapy decisions for diabetic patients. In this approach, the blood glucose concentration is continuously monitored through small wearable devices/systems, which track the glucose concentration levels in the blood round-the-clock. Such systems could be invasive, minimal-invasive, or non-invasive. Moreover, the CGM systems can be classified into two types: retrospective systems and real-time systems [10].

The introduction and availability of a variety of innovative CGM devices/systems open new opportunities for diabetic patients to manage glycemic control with ease. Most of the modern CGM devices normally compute and record the current glycemic state of a patient every minute through continuous measurement of interstitial fluid (ISF) by adopting a minimally invasive mechanism. These systems/devices are minimally invasive, since they compromise the skin barrier but do not puncture any blood vessels. Besides, there are non-invasive methods, for instance, measuring blood glucose concentration by applying electromagnetic radiation through the skin to the blood vessels in the body [11]. Moreover, the emergence of e-Health in the form of telemedicine not only enables the physicians to observe the patients remotely and regularly, but also transmits the CGM data to the remote database in the hospital, which could be used to forecast critical events in the glycemic control such as hypo/hyperglycemia.

One of the challenges in diabetes management is the prevention of hypo/hyperglycemia events, which could be overcome by accurately forecasting the blood glucose concentration from the CGM/SMBG and related (i.e., exercise, food intake, insulin intake, etc.) data. Thus, the development of tools for the processing and interpretation of CGM/SMBG and diabetes related data for future glucose values is crucial. To this end, data mining plays an important role in the development of such tools for the diagnosis and prediction of diabetes [12], [13]. Data mining is a process of extracting valuable information from a large volume of data in order to discover previously unknown trends, patterns, and relationships that could be used to build models for prediction [14]. In the literature, different data mining based glucose forecasting approaches and methods have been developed based on various models. These techniques extract, analyze, and interpret the available diabetes data in order to make clinical decisions. A generic framework of such techniques is shown in Figure 1.

In this paper, we present a state-of-the-art review in the field of glycemic control concerning the diabetes diagnosis and prediction using data mining. We classify the commonly used data mining based solutions for diabetes diagnosis and prediction based on the underlying model used. Moreover, we compare them based on key parameters and metrics. Finally, we point out the challenges that need to be addressed and future research directions in the area.

The remainder of the paper is organized as follows: Section II provides a thorough discussion on the related work. The methodology of the survey is given in Section III. Section IV presents the classification and comparison of data mining based diabetes diagnosis and prediction techniques. The challenging open issues and future directions are discussed in Section V. Lastly, Section VI concludes the paper.

II. RELATED WORK

A diverse literature has contributed to the area of diabetes diagnosis and prediction ranging from the development and performance analysis of novel data mining based techniques for diabetes detection, prediction, and classification, to the survey and review studies, as can be seen in [15]–[25]. In [16], [24], [26], various data mining techniques for diabetes detection are reviewed and discussed. Similarly, in [17], a systematic review of the application of data mining techniques for diabetes, as well as the corresponding data sets, methods, software, and technologies, is carried out. Based on this review, it is concluded that data mining has a key role and bright research future in the field of glycemic control. Data mining is used to extract valuable information from diabetes data, which ultimately helps diabetic patients in the

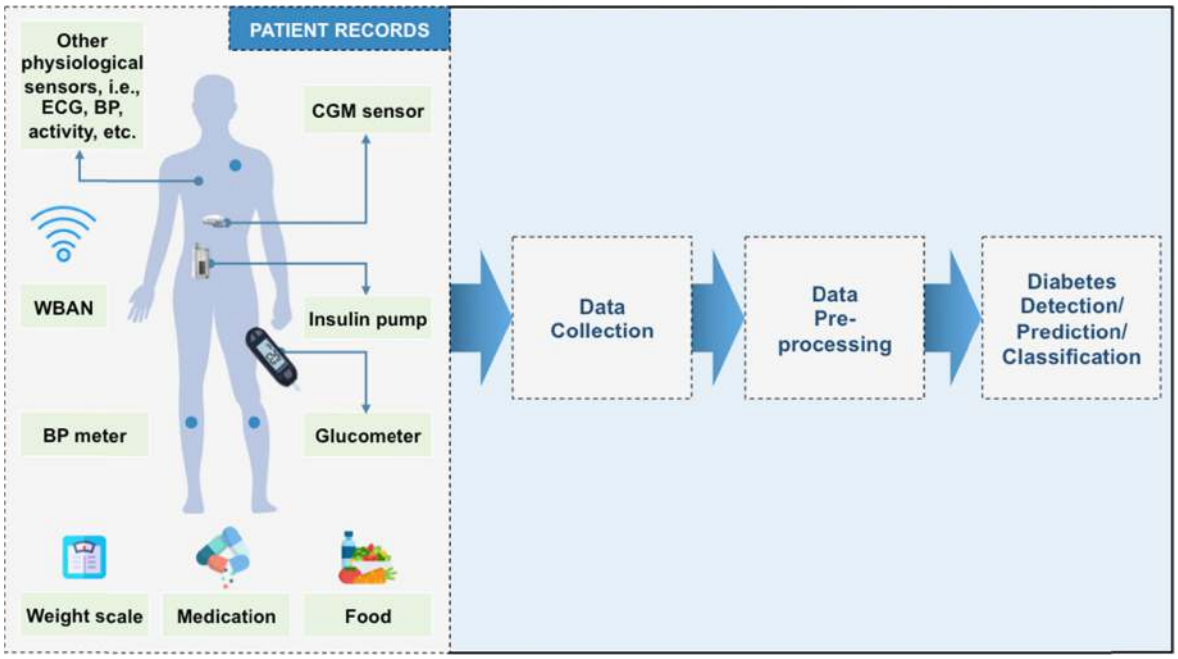

FIGURE 1. A generic flow of diabetes detection and prediction techniques.

management of their glycemic control. Likewise, in [18], a survey is conducted on the application of different data mining techniques, including artificial neural network (ANN), for the prediction and classification of diabetes. The survey shows that ANN outperforms the rest of the techniques with 89% of prediction accuracy.

On the other hand, in [19], the performance of four well-known methods, namely J48 decision tree (DT) classifier, KNN, random forests algorithm, and support vector machine (SVM), is evaluated in terms of prediction of diabetes using data samples with and without noise from the University of California Irvine (UCI) machine learning data repository [27]. From the comparative analysis of these techniques, it is observed that J48 classifier performs better in the presence of noise in the data with 73.82% accuracy. Whereas in case of noise-free data, the KNN (k=1) and random forests outperform the rest of the two methods with an accuracy of 100%. Furthermore, in [21], with the help of data mining tools such as WEKA, TANAGRA, and MATLAB, a comparative study of nine different techniques is performed in the light of diabetes prediction using pima Indian diabetes dataset (PIDD) from UCI machine learning repository [27]. According to the performance analysis, the best classifiers in WEKA, TANAGRA, and MATLAB are J48graft, NB and adaptive neuro-fuzzy inference system (ANFIS) with the corresponding accuracies of 81.33%, 100%, and 78.79%, respectively. Likewise, in [20], [23], [28], the comparison and performance evaluation of various data mining techniques are presented.

In [29], a study is conducted based on six diabetes intervention models using SVM classification technique. The comparative analysis shows that smoking cessation is the best intervention with high accuracy. Moreover, in [30], a method based on data driven model is proposed for the glucose prediction using a multi-parametric set of free-living data such as food, activity, and CGM data. In this method, the effect of diet, physical activity, and medication on the glucose control is investigated. The method incorporates the meal model, exercise model, insulin model, and glucose prediction model based on support vector regression (SVR). The evaluation on data (CGM, activity insulin, etc.) from seven type 1 diabetic patients shows promising results for 15 and 30 minutes of predictions.

Furthermore, feature selection, extraction and classification, and dimensionality reduction play an important role in the prediction of risk events in glycemic control. In the literature, abundant work has been presented on the feature extraction and classification, as shown in [31]. In [32], a hybrid prediction model is constructed. In order to improve the prediction accuracy, the model is evaluated using two types of data from the PIDD [27]: data without feature selection and data with feature selection. Based on a comparative analysis from these two scenarios, it is observed that the overall detection accuracy improves with feature selection. Similarly, in [33], a method is proposed for the diagnosis of diabetes based on bi-level dimensionality reduction and classification algorithms using PIDD [27]. The bi-level dimensionality reduction includes feature selection for removing irrelevant features and feature extraction. The diabetes data analysis with bi-level dimensionality reduction using different data mining techniques shows increased performance.

Based on our thorough literature review, we observe that most of the existing research either discusses the evaluation

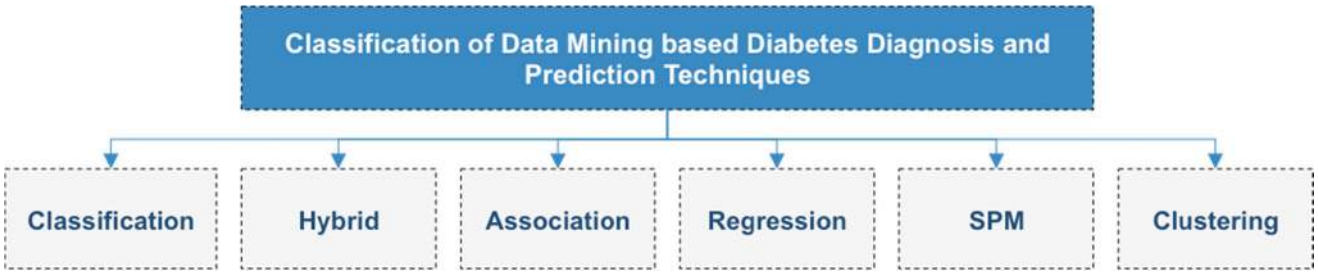

FIGURE 2. Broad classification of diabetes diagnosis and prediction approaches.

of existing data mining based diabetes detection, prediction, and classification techniques, or present brief surveys on few of such techniques. However, to the best of our knowledge, none of these covers a comprehensive classification and comparison of the existing techniques and the corresponding challenging issues in this domain. In order to provide a comprehensive classification and comparison of existing techniques using key parameters and to highlight the corresponding challenges in the field of diabetes detection, prediction, and classification based on data mining models, in this work, we present a comprehensive state-of-the-art survey on the development of overall systems for diabetes diagnosis and prediction. Moreover, the corresponding challenges are discussed and various open issues are highlighted for future research in the field of glycemic control.

III. SURVEY METHODOLOGY

Aiming to conduct our survey using a systematic approach, and in order to delimit the theme of the survey, it was necessary to elaborate inclusion and exclusion factors to define which aspects would be valued and which would not be considered in the survey. This section describes the aspects that formed the documentary basis of this article.

A systematic search of well-known bibliographic scientific databases including Clarivate Web of Science (WoS), PubMed, and Google Scholar was conducted by using related key words such as ''Machine learning'', ''Data mining'', ''Diabetes mellitus'', ''type 1 diabetes'', ''type 2 diabetes'', ''Gestational Diabetes Mellitus'', ''Prediction'', ''Detection'', etc. Then, the inclusion and exclusion factors were elaborated so that we could obtain a significant amount of studies to be analyzed. The abstracts of the searched papers were studied and evaluated in detail to check the eligibility of the research for inclusion. The original studies related to diabetes in which data mining techniques were used for diagnosis and prediction of the disease, were considered appropriate for inclusion. After selecting the studies, full research articles were downloaded. Secondary reports such as editorials, news articles, brief communications, and non-peer-reviewed correspondence/articles were not included. We reviewed a large number of papers for diagnosis and prediction of diabetes using data mining techniques, out of which 80 studies were selected that belonged to a specific class. A total of six classes/categories were identified based on the underlying data-mining model used. Publication years of selected works range from 2006 to 2020. We included research works published in peer-reviewed journals and in some highly cited and well-known related conferences.

IV. CLASSIFICATION OF DATA MINING BASED TECHNIQUES FOR DIABETES DIAGNOSIS AND PREDICTION

The glucose concentration in diabetics is influenced by various factors, such as meal intake, physical activity, drug intake, insulin, emotions, stress, etc. As a result, the glucose concentration varies with time, which can lead to risky events, such as hypo/hyperglycemia, if it is not properly managed. Therefore, the prediction of future glucose concentrations in diabetics is an important, interesting, and challenging research area yet to be fully explored by the research community. To this end, in order to predict diabetes and cope with hypo/hyperglycemia, numerous predictive and prescriptive mining approaches have been developed for the forecasting of glucose concentrations and detection of such events. In this section, we present our classification of such techniques based on the underlying data mining models used, as shown in Figure 2. Table 1 describes the symbols for different parameters used in Tables 2-7. Under each category, various recent techniques are compared using key metrics, as presented in Tables 2-7, and are discussed in subsequent sections accordingly. A brief description of each class along with the corresponding schemes is given below.

A. CLASSIFICATION-BASED TECHNIQUES

Classification is a supervised learning process in which a class of objects is classified in order to predict any classes of future objects [34]. In the literature, numerous classification based diabetes prediction techniques have been developed [3], [35]–[56]. In [56], authors proposed a random forests classifier with the genetic algorithm. The goal of the classifier is to assist in medical diagnosis by extracting the required information from the symptoms exhibited by a patient. A set of experiments was done to compare the proposed approach with other hybrid classifiers for diabetes mellitus and it was found that the approach outperformed other algorithms in the metrics used. It had an accuracy of 0.923, sensitivity of 0.901,

TABLE 1. Symbols and descriptions.

Symbol	Description	Symbol	Description	Symbol	Description
Alg/Mod	Algorithm/Model	KNN	K-nearest Neighbor	NB	Naive Bayesian
Data Prep	Data Preprocessing	DT	Decision Tree	RMAR	Regular & Maximal Association Rules
N/A	Not Applicable/Not present/Not Discussed	AA	Apriority Algorithm	AR	Autoregressive
Imp	Implementation	SVR	Support Vector Regression	RPLS	Recursive & Partial Least Squares
Pnp	Plug-n-play	BT	Bayes Theory	LS	Least Squares
Sim	Simulation	SAR	Survival Association Rule	HC	Hierarchical Clustering
MDC	Model-based Clustering	PBC	Partitioning-based Clustering	DBSCAN	Density-Based Clustering
PLWAP	Pre-Order Linked Web Access Pattern	MST	Minimum Spanning Tree	SVM	Support Vector Machine
ANN	Artificial Neural Network	QDA	Quadratic Discriminant Analysis	IBL	Instance Based Learner Classifier
RF	Random Forest	XPRESS	eXtraction of Phenotypes from Records using Silver Standards model	HCA	Hierarchical Clustering Algorithm
GLM	Generalized Linear Model	PR	Penalized Regression	LDA	Linear Discriminant Analysis
SQ-MRM	Semi-Quantitative Multiple Reaction Monitoring	SID-MRM	Subsequent stable-Isotope Dilution Multiple Reaction Monitoring	BR	Bootstrap Resampling
MAM	Meta-Analytic Methods	RE-MAM	Random-Effects Meta-Analysis Methods	FWA	Filter and Wrapper Approaches with P-value-based methods
CSS	Chi-Square Statistic	IG	Information Gain	GR	Gain Ratio
SU	Symmetrical Uncertainty Criterion Compensates	NHC-MV	Novel Hierarchical Clustering based on Minimum Variance	RAM	Regression Activation Maps
CNN	Convolutional Neural Network				

specificity of 0.924, and Kappa Statistics of 0.879. In terms of future work, the authors proposed research and development towards blending the algorithm with hybrid genetic algorithms, a step aimed at improving the performance of the approach even further.

In [57], authors looked into developing a data analysis approach whereby gases and volatile organic compounds (VOCs) were measured using non-invasive samples with a field asymmetric ion mobility spectrometry (FAIMS) approach. The work affirmed that processing with a 2D wavelet transform is a preferred option than using a 1D wavelet transform. The experiments were done in a 2-step feature selection process with the first step filtering out low variance features. This was then followed by a step where the information features were selected using a filter method known as the Wilcoxon rank-sum test. The first step was found to have less impact in the process but the latter added to the quality of the process by minimizing dimensionality of the data and improving the AUC scores. The filter approach used in the second step also reduced the computation time and the prediction metrics of the classifier. The authors also experimented with the idea of adding principal component analysis (PCA) in the data analysis pipeline. The goal of adding PCA was to filter out the effect of unrelated features but it was found to have a negative effect on the AUC scores. The authors concluded that using linear combinations of the features selected might have a negative effect on the signals in which they were interested.

In [35], an online method is developed for the future predictions of interstitial glucose concentration levels from the CGM data, where an ANN model is used for the implementation of the predictor. The model takes the CGM sensor values of the past 20 minutes as an input and provides the prediction of the glucose concentration as an output at the selected prediction horizon (PH) time. The presented scheme showed better prediction accuracy for different PHs, i.e., 15, 30, and 45 minutes, with more accuracy, and no significant deterioration in the prediction delay compared to that of an AR model based scheme in [3]. Nevertheless, the proposed scheme would not be able to detect sudden glucose variations due to meal intake, insulin intake, and physical activity, etc., as it only depends on the CGM data. Besides, the scheme is CGM systems dependent and is not a generic one.

In [38], an ANN model based glucose levels prediction method is proposed for the prevention of the hypo/hyperglycemia events in critically ill trauma patients admitted to the hospitals. In this method, the aim is to develop and optimize patient-specific and general ANN models that could provide real-time prediction of glucose concentrations in critically ill patients in 75 minutes of PH. The method is evaluated with acceptable results in terms of prediction; yet, the method is not implemented in real-time.

TABLE 2. Comparison of diabetes diagnosis and prediction techniques based on classification class.

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
[35]	ANN	CGM data	1, 2	N/A	15 type 1 diabetics data	Sim	No	Prediction of future glucose concentration levels from CGM data
[37]	ANN	CGM data, food intake, insulin intake	1	N/A	4 type 1 diabetics data	Sim	No	Modeling glucose-insulin metabolism for children with type-1 diabetes
[54]	ANN	CGM data, meal and ingested carbohydrates	1	N/A	20 type 1 diabetics	Sim	No	Prediction of blood glucose level
[53]	ANN	CGM data, meal	1	N/A	9 real datasets and 20 simulated datasets	Sim	No	Short-time prediction of glucose concentration
[46]	ANN	CGM data	1	N/A	Data from JDRF CGM study group	Sim	No	Prediction of Hypoglycemia Concentration
[38]	ANN	CGM data, medical records (glucose test times, insulin)	1, 2	N/A	Data obtained from a 38-year old trauma patient	Sim	No	Prediction of blood glucose level
[39] [40]	ANN	CGM data, SMBG data, insulin dosages, carbohydrate intake, hyperglycemic and hypoglycemic symptoms, lifestyle (activities and events), and emotional states	1	N/A	Data from 18 insulin-dependent patients and Data from 27 patients	Sim	No	Prediction of blood glucose level
[41]	ANN	CGM data	1	Yes	9 Type 1 diabetics	Sim	No	Prediction of blood glucose level
[42]	ANN	SMBG data, insulin, food, stress, exercise	1, 2	Yes	Data of continuous period of 77 days from 1 patient	Sim	No	Prediction of blood glucose level
[43]	ANN	Simulated glucose values, meal and insulin intake	1	N/A	28 datasets (from a single case scenario) from AIDA simulator	Sim	No	Prediction of blood glucose level
[55]	ANN	SMBG data, carbohydrates, exercise, insulin	1	N/A	Data from 6 patients	Sim	No	Prediction of the amount of insulin to inject
[36, 49]	SVR	CGM data, food intake, physical activity, insulin injection	1	N/A	3 Type 1 diabetics	Sim	No	Prediction of blood glucose level
[51]	SVR	Breath samples	1, 2	Yes	295 healthy and 279 diabetes breath samples	Sim	No	Prediction of blood glucose level
[52]	SVR	CGM data, meal, insulin	1	Yes	5 type 1 diabetics	Sim	No	Prediction of blood glucose level
[50]	SVR	CGM data, meal intake, insulin intake, exercise	1	Yes	Data from 4 diabetes support system (4DSS)	Sim	No	Prediction of blood glucose level and decision support for diabetes

TABLE 2. *(Continued.)* **Comparison of diabetes diagnosis and prediction techniques based on classification class.**

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
								management
[47]	DT	Plasma glucose, BMI, DPF, age, BP, pregnancy	2	Yes	Pima Indians Diabetes Data Set	Sim	No	Prediction of patients with developing diabetes
[48]	DT	CGM data	1	Yes	Data from 10 patients	Sim	No	Predicting hypoglycemia in diabetic patients
[45]	BT	CGM data, meal, insulin	1	Yes	20 simulated datasets and clinical dataset	Sim	No	Predicting glucose in insulin-dependent diabetes
[59]	SVM, DR	The Messidor database of eye fundus color numerical images, Blood vessels, Microaneurysms, Hard exudates	1, 2	N/A	400 retinal images labeled with respect to a 4-grade scale of nonproliferative diabetic retinopathy.	Sim	No	Detection of diabetic retinopathy
[12]	NB, SVM, DT	Pregnancy, Diastolic blood pressure, Plasma glucose, Skin fold thickness, Diabetes pedigree function, Serum insulin, BMI, Age	1	N/A	768 women patients' data from PIDD dataset	Sim	No	Diabetes prediction
[60]	NB, QDA, LR, IBL, SVM, ANN, RF	Diverse set of attributes	1, 2	Yes	5 heart disease datasets, 2 diabetes datasets, 4 breast cancer datasets, 2 liver disease datasets, and 1 hepatitis dataset taken from public repositories.	Sim	No	Diabetes prediction
[44]	ANN, SVR, NB	Annotated CGM plots data	1, 2	Yes	CGM annotated plot dataset of 218 plots	Sim	No	Detection of excessive glycemic variability

Authors in [58] were interested in developing an algorithm for predicting diabetes on the basis of data mining approaches, such as clustering and classification, with the objective being to diagnose the disease early; hence allowing timely and appropriate treatment. The authors found classification to be the better technique with J48 and Naïve Bayesian approaches as the most suitable. The authors then proposed the model and presented an intelligent diagnostic system that makes it easy for the practitioner to comprehend the discovered rules. Experiments were done based on a dataset collected from an online portal and a college medical hospital, but clinical trials and other research work were still ongoing.

On the other hand, in [39], the predictive feasibility of NN models in different predictive windows, i.e., from 50 – 180 minutes, for glucose concentrations in diabetes patients, is presented. In this study, the ANN models are trained with patients' CGM data along with data related to drugs intake, meal intake, SMBG values, hypo/hyperglycemic symptoms, life events, and emotional states recorded in a PC-based electronic diary by the patients. The model's performance is acceptable in terms of hyperglycemia and normal concentration prediction. Nevertheless, they overestimated the hypoglycemia, which according to the authors was due to the less numbers of hypoglycemic events in the underlying training data. This method involves manual data inputs in the diary and manual integration of CGM values in electronic diary that can incorporate errors in the modeling; and hence, in the prediction. In another work [40], the authors implemented and assessed the ANN model for real-time prediction of glucose concentration for a PH of 75 minutes. In this particular work, the design of ANN model is different from the designs of the complex models used in [39]. In [39], the ANN models are based on time-lagged feed-forward neural network and the memory element to store the inputs' historical values for the quantification of trends and patterns in the historical data. Whereas in [40], a relatively simple feed-forward model design is adopted, which comparatively reduces the complexity of the architecture and the required processing time for real-time training and prediction. In addition, the model is evaluated on a relatively large dataset and used RMSE and CEGA metrics along with MAD for the performance analysis. Likewise, in [41], a glucose prediction model is developed for type 1 diabetics using CGM values in a feed forward neural network. The authors

TABLE 3. Comparison of diabetes diagnosis and prediction techniques based on regression class.

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
[3]	AR	CGM data	1	Yes	28 Type 1 diabetics	Sim	No	Prediction of blood glucose level
[64]	AR	CGM data	1, 2	Yes	Data from 34 subjects with either type 1 or type 2 diabetes	Sim	No	Prediction of blood subcutaneous glucose level
[65]	AR	CGM data	1	Yes	9 type 1 diabetics data collected over a 5-day period	Sim	No	Prediction of blood subcutaneous glucose level
[67]	AR	CGM data	1, 2	Yes	27 Type 1 diabetics and 7 type 2 diabetics	Sim	No	Prediction of blood subcutaneous glucose level
[66]	RPLS AR	CGM Data	1	Yes	Real data of 17 subjects and simulated data of 20 subjects	Sim	No	Prediction of blood glucose level
[68]	GARCH (1, 1)	CGM data	1	N/A	Real data of 6 patients	Sim	No	Predicting Glucose Levels in Patients with type-1 Diabetes
[69]	ARX	CGM data, insulin intake, meal intake	1	N/A	Data from 4 type 1 patients	Sim	No	Online prediction of blood glucose profile in type-1 diabetes patients
[70, 71]	AR, ARIMA (3,1), ARIMA (2,1)	CGM data	1, 2	Yes	25 type 2 diabetics CGM data and simulated Type 1 diabetes data	Sim	No	Estimation of future glucose level
[72, 73]	AR, ARX, LV	CGM Data, meal CHO estimate and bolus insulin	1	N/A	17 type 1 diabetics data	Sim	No	Online prediction of subcutaneous glucose concentration for type 1 diabetes
[77]	XPRESS	Diagnosis data of patients with T2DM and MI	2	Yes	Data collected by The Stanford clinical data warehouse (SCDW) for 1.2 million patients over 19 years	Sim	No	Predicting Type 2 diabetes mellitus and myocardial infarction
[78]	HCA, GLM, PR	high-resolution genotypes of HLA genes	1	N/A	962 cases and 448 controls for training and validation	Sim	No	Predicting type 1 diabetes with class II HLA genes
[79]	ANN, LR, SVM,	Gender, Age, Alcoholic cirrhosis, Other cirrhosis, Alcoholic hepatitis, Viral hepatitis, Other chronic hepatitis, Alcoholic fatty liver disease, Other fatty liver disease, Hyperlipidemia	2	Yes	1442 cases divided into training group (70%) and a test group	Sim	No	Predicting liver cancer for type II diabetes patients
[80]	Binary LR, NB	No. of subjects with and without type 2 diabetes, No. of subjects with a HW phenotype, Age, TG, FPG, Systolic blood pressure, Diastolic blood pressure, Weight, Body mass index, Circumference of (Neck, Chest, Rib, Waist, Hip, NeCk-to-hip ratio, Rib-tO-hip ratio, Waist-to-hip ratio, Forehead-to-waist ratio, Forehead-to-rib ratio, Forehead-to-neck ratio,	2	Yes	Data collected by Korean Health and Genome Epidemiology Study database from 11,937 subjects (4,906 males and 7,031 females) over 7 years	Sim	No	Identification of type 2 diabetes risk factors

TABLE 3. *(Continued.)* **Comparison of diabetes diagnosis and prediction techniques based on regression class.**

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
		Waist-to-height ratio).						
[81]	LDA, SQ-MRM, SID-MRM	Plasma and protein samples attributes	2	Yes	96 proteins were and selected from 601 collected from 20 diabetic patients.	Sim	No	Early stages detection for diabetes
[82]	BR, MAM,	Age, systolic blood pressure, respiratory rate, low sodium serum concentration, serum urea nitrogen, low serum hemoglobin, presence of dementia, presence of cerebrovascular disease, hepatic cirrhosis, chronic obstructive pulmonary disease, and cancer	N/A	N/A	Data collected from 14,857 patients at 90 hospitals in two distinct time periods	Sim	No	Validation of clinical prediction models
[83]	RF, FWA, CSS, IG, GR, The SU, ReliefF	Variables obtained from blood samples	2	Yes	Data collected from 803 prediabetic females	Sim	No	Predicting of diabetes
[84]	Phenotyping algorithms	hyperlipidemia (HLD) to hypertension (HTN), impaired fasting glucose (IFG), and type 2 diabetes mellitus (T2DM)	2	Yes	Data collected from 70k patients over 13 years from Mayo Clinic	Sim	No	Identifying type 2 diabetes mellitus trajectories
[85]	KNN, DT, NB	Demographics, socio-economic, BGL, HbA1c, cholesterol profile, inflammatory and oxidative stress, medical history, BMI, peripheral vascular function, and ECG	2	Yes	Data collected from 840 patients over 10 years by Diabetes Health screening clinic (DiabHealth)	Sim	No	Identifying type 2 diabetes mellitus trajectories
[74, 75]	ARX, LS	CGM, insulin, meal	1	Yes	30 type 1 diabetics simulated data	Sim	No	Identifying glucose concentration prediction

train three ANN models with 15, 30, and 45 minutes of PH respectively. The analysis results show that the proposed model can predict accurately compared to other such models with no significant time delay in the predicted values and the real glucose values in the PH of 30 minutes or less. Moreover, according to the authors, the method is fully automated and does not require any inputs from the patient other than the automated CGM values from CGM sensors on the patient's body [61]. It only considers CGM data, which might not be sufficient for accurate forecasting if the effects of drugs and other life events on the glucose concentrations are considered.

In [25], authors performed a systematic review on the use of machine learning and data mining technology in research on various diabetes areas. These fields include biomarker prediction and diagnosis, diabetic complications, genetic background and environment, and healthcare management. The emphasis was much laid on the first field and several machine learning algorithms were used. 85% of those involved supervised learning approaches, while the remaining 15% were characterized by unsupervised learning approaches and association rules. From the analysis, the SVM emerged as the most successful and widely implemented algorithm.

A computer-aided diagnosis model is proposed in [59] to perform digital processing of retinal images so as to facilitate the early detection of diabetic retinopathy. Authors were interested in grading non-proliferative diabetic retinopathy at a given retinal image. First, the blood vessels, microaneurysms, and hard exudate are isolated to enable the extraction of necessary features for vector machine to grade each retinal image. Then, this model is examined using 400 samples of retinal images labeled as per the 4-grade scale of non-proliferative diabetic retinopathy. The experimental results indicated that the proposed method gathered a maximum sensitivity of 95% and predictive capacity of 94%.

A predictive model is proposed in [12] to evaluate the possibility of diabetes in patients with high accuracy levels. A number of machine learning algorithms are used in the experimental phase, i.e., Decision Tree, SVM, and Naive Bayes. The PIDD dataset from the UCI machine learning repository is used. Three machine learning algorithms are then evaluated on the basis of a number of measures. Specifically, the measures of interests in the experiment

TABLE 4. Comparison of diabetes diagnosis and prediction techniques based on association class.

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
[88]	AA	Demographic characteristics, anthropometric measures, biochemical measures, disease history, medical history	2	Yes	6647 non diabetics' people data	Sim	N/A	Extracting risk pattern for type 2 diabetes
[89]	AA	SNPs data	2	Yes	Type 2 diabetes dataset of 92 SNPs	Sim	No	Analyzing genetic data of type 2 diabetes
[90]	SAR	Demographic, systolic blood pressure, diastolic blood pressure, total cholesterol, high-density lipoprotein, low-density lipoprotein, BMI, triglycerides, diagnoses, medication	2	N/A	Clinical data set of 21,981 pre-diabetic patients	Sim	No	Risk assessment of type 2 diabetes
[91]	AA	Gender, age, occupation, complications, medical treatment method of expense	2	Yes	Dataset of 3,964 patients with ophthalmic complication	Sim	No	Risk assessment of type 2 diabetes
[92]	AA	Diagnosis data of patients with T2DM	2	Yes	Dataset of 411,414 patients' clinical record	Sim	No	Finding association rules of diabetes mellitus with ophthalmic complication
[93]	SAR	Co-morbid diseases, laboratory results, medications and demographic information	2	N/A	Medical record of 23,828 patients	Sim	No	Analyzing comorbidity in patients with type 2 diabetes mellitus (T2DM)

are Precision, Accuracy, F-Measure, and Recall. The Naïve Bayes algorithm achieved better performance than the other two algorithms, a fact that was confirmed by receiver operating characteristic (ROC) curves during the analysis.

Furthermore, a multi-layer framework for classification, known as HM-BagMoov is proposed in [60]. Essentially, this is an ensemble of 7 classifiers containing bagging and optimized weighting targeted at overcoming the issues associated with the conventional classifiers. Authors evaluated the framework on 5 heart disease datasets, 4 breast cancer datasets, 2 diabetes datasets, 2 liver disease datasets, and 1 hepatitis dataset, all from publicly available repositories. On experimenting with the framework, it was found that the approach had high accuracy, sensitivity, and F-measure compared to other common approaches. Consequently, the authors went on to develop an app by the name IntelliHealth on the basis of the proposed model, which can be used by healthcare professionals to assist with diagnosis guidance.

Authors in [62] introduced a novel method that leveraged data mining algorithms in the prediction of type 2 diabetes mellitus (T2DM). The main objective of the work was to aid in improving accuracy when it comes to the prediction model, whilst also making it adaptable to different types of datasets. The model has a number of preprocessing steps and two main parts, which are the enhanced K-means algorithm and the logistic regression algorithm. The PIDD dataset and the Waikato Environment for Knowledge Analysis (WEKA) toolkit were used for the experimental setup, where the proposed method was compared to others. The results showed that the data mining approach had better accuracy compared to methods proposed by other researchers. To affirm its prediction ability, the authors also experimented on two additional diabetes datasets. The experiments also showed good performance levels, thus indicating that the prediction algorithm could actually be proven useful in the health management of diabetes.

Moreover, in [42], a blood glucose level predicting system is presented, which is based on the integration of PCA and wavelet neural network (WNN) with different wavelet families such as Morlet, Mexican Hat, and Gaussian wavelet embedded in its hidden layer. The PCA extracts features from the data, which are then used in the WNN models for prediction in different intervals including morning, afternoon, evening, and night. This method not only uses data from one patient to train the models, but it also relies on the data such as meal, etc., which is provided by the patient manually. Systems developed under such conditions might not produce concrete results. In contrast, in [43], a glucoregulatory model is formulated based on Elman recurrent ANN using data generated from automated insulin dosage advisor (AIDA) freeware simulator. The proposed method shows good results for short-term and long-term prediction during night times of the daily cycle. However, the method is based on a single subject data, which might not be sufficient

TABLE 5. Comparison of diabetes diagnosis and prediction techniques based on clustering class.

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
[97]	HC	Lifestyle, symptoms	All	N/A	Diabetes data collected from different medical websites (webmed, Mayo clinic, etc.,)	Sim	No	Diagnosis and prevention of diabetes
[98]	MDC	Socioeconomic, demographic, and environmental characteristics	2	N/A	Data collected from the US states of Pennsylvania and New York	Sim	No	Prediction of rates of obesity and diabetes
[99]	HC	Demographic, laboratory, diagnosis, aspirin, medications and the use of tobacco information	2	Yes	52,139 patients' data from Mayo Clinic	Sim	No	Identification of clinically relevant patient sub-populations using type 2 diabetes
[100]	PBC	Laboratory, diagnosis, age, number of pregnancies	2	Yes	268 diabetic women data from National Institute of Diabetes, Digestive and Kidney Diseases	Sim	No	Predicting initial or advanced stage of diabetes
[101]	DBP	Patients' clinical examination history data	All	Yes	6,380 patients' examination Data from National Health Center (NHC) of the Asti province (Italy)	Sim	No	Identifying examination pathways followed by diabetic patients
[102, 103]	HC	Patient medical record such as blood sugar, fat, cholesterol, or potassium	2	N/A	27 days blood sugar records of a single anonymous patient	Lab Test	No	Analysis and Prediction of diabetes patients' conditions
[104, 105]	HC	Patient's medical data such as BMI, BP, glucose level, cholesterol, waist circumference, triglyceride	2	N/A	2 patients' examination data of over 9 years	Sim	No	Identifying temporal progress of metabolic syndrome patients' health status
[106]	NHC-MV	Gender, age, race, diagnostic information, blood test results	1, 2	Yes	Data collected from 3041 patients over 4 years.	Sim	No	Probability prediction of readmission in diabetic patients
[107]	ANN	General condition data, Heart condition, cancer, brain disease/ neurodegenerative disease, mental disorder, substance abuse questions	2	N/A	Collected information from 2,412 individuals (709 children, 1,703 male and female)	Sim	No	Predicting the likelihood of diabetes and other diseases
[108]	LR, SVM and ANN	pH and oxidation reduction potential (ORP) values, conductivity and concentration of the electrolyte, FBGL	2	Yes	Data from 175 individuals in the age range of 18–69 years.	Sim	No	Detection of fasting blood glucose levels
[109]	RAM, CNN	Diabetic retinopathy for right and left eyes	1,2	Yes	35126 high resolution labeled images of diabetic retinopathy	Sim	No	Diabetic retinopathy detection
[86]	DT, LR, NB	Fatty acids, amino acids, biocraters, ketone body, glucose, age, pre-pregnancy BMI, race/ ethnicity	2	No	Data collected from 1,035 women with GDM pregnancy	Sim	No	Predicting the transition from GDM to type 2 diabetes

for accurate results. On the other hand, the data itself is from a source that is already declared unreliable for diabetes therapy planning and contains many glitches; therefore, results on such data might be uncertain. Furthermore, in [37], the combination of compartmental models (CMs) and ANN models are used for simulating glucose-insulin metabolism of children for the purpose of short-term prediction of glucose concentrations in type 1 diabetics. In this scheme, the CMs estimate the influence of food intake on the blood glucose and the effect of insulin intake on the insulin concentration of plasma, which are then fed to the ANN model along with previous glucose values for prediction. Moreover, for

TABLE 6. Comparison of diabetes diagnosis and prediction techniques based on SPM class.

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
[117, 118]	AA	Insulin dosage, blood glucose measurements	1	Yes	Dataset of 102 children type 1 diabetics from Silesian Medical University in Katowice, Poland	Sim	No	Providing insulin therapy recommendations for type-1 diabetes
[119]	PLWAP Tree	Glucose measurements, insulin, meal, exercise, hypoglycemic symptoms	2	Yes	Diabetes dataset from UCI machine learning data repository, Washington university	Sim	No	Guiding diabetic patients to manage their health
[120]	CSPADE	Drug class, generic drug name only, generic drug and dose, or brand name and dose	2	Yes	161,497 patients data from Blue Cross Blue Shield of Texas for patients	Sim	No	Predicting next medications to be prescribed for diabetic patients

TABLE 7. Comparison of diabetes diagnosis and prediction techniques based on hybrid class.

Scheme	Alg/Mod	Data Input	Type	Data Prep	Evaluation Dataset	Imp	Pnp	Application
[121, 122]	SVM , RBF, CFP-Growth++	Medical records of 768 patients	2	Yes	Pima Indian Diabetes Dataset at UC Irvine Machine Learning Lab	Sim	No	Prediction and early detection of type-2 diabetes
[123]	HC, MST, KNN	SNP (genotype data)	2	N/A	1999 case and 1999 control individuals genotype data on Type 2 diabetes from Wellcome Trust Case Control Consortium (WTCCC)	Sim	No	Type-2 diabetes susceptibility prediction
[124]	SVM, RTP	Laboratory tests and the diagnosis codes (related diseases)	1, 2	Yes	13558 record of adult diabetic patients	Sim	No	Detection of adverse medical conditions associated with diabetes
[126, 125]	DT, ANN / DT, PBC	Plasma–glucose, BP, triceps skin fold thickness, serum insulin, BMI, age, frequency, status of diabetes (yes/no), diabetes pedigree function	GDM, 2	Yes	Dataset unspecified / 768 randomly selected female patients from Pima Indians diabetes dataset obtained from UCI machine learning repository	Sim	No	GDM prediction/ Prediction for type-2 diabetes
[127]	PBC, SVM	Pregnancy frequency, plasma–glucose, diastolic BP, triceps skin fold thickness, serum insulin, BMI, diabetes pedigree function, age, class variable	2	Yes	Pima Indians diabetes dataset from UCI machine learning repository	Sim	No	Diagnosis of diabetes
[128]	HC, SVM	Patient profiles, hospital profiles, diagnostic profiles and procedure profiles	2	N/A	Healthcare Cost & Utilization Project (HCUP-3) database	Sim	No	Detection of type-2 diabetes
[129]	AR, SV, ELM	CGM data	1	N/A	10 type 1 diabetics from JDRF CGM Study Group	Sim	No	Blood glucose prediction
[130]	cARX, RNN	CGM data, insulin pump data, hypo/hyper glycemic events	1	N/A	23 Type 1 diabetics	Sim	No	Early warning of future hypoglycemic/hyperglycemic events for type-1 diabetes

comparison purposes, real-time recurrent learning algorithm trained recurrent neural network (RNN) and feed-forward neural network (FFNN) architectures are evaluated with better prediction performance achieved by the former. However, the proposed scheme can be further improved by incorporating physical activity data in the analysis. Additionally,

the proposed scheme is evaluated on limited subjects, which reduces the significance of the results. Conversely, according to [44], excessive fluctuation of blood glucose concentration levels could also lead to diabetic complications. Therefore, in [44], the authors have developed an approach based on multilayer perceptron (MP), NB and SVM classifiers for automatic detection of excessive glycemic variability in preprocessed CGM data. This method could be used as an automated screen for the detection of glucose concentrations variability in diabetics based on the physician annotated CGM data graphs. Besides, in [45], an ensemble method, merging different predictor models, is proposed for future glucose prediction, which incorporates a Bayesian framework. The method is named as sliding window Bayesian model averaging (SW-BMA) predictor. Before the analysis, this method performs feature extraction. The proposed technique is evaluated against three individual predictors both on simulated and empirical datasets. The SW-BMA shows improvement in performance compared to individual predictors. Additionally, in [46], for the prediction of hypoglycemia, extreme learning machine (ELM) and regularized ELM methods are used. These two methods are evaluated over a real CGM data for PH of 10, 20 and 30 minutes with comparable results. The methods lack the ability to detect sudden glucose changes due to metabolism since it does not consider other physiological data.

On the other hand, in [47], a decision tree model based method is developed for the prediction and classification of patients with developing type 2 diabetes. In this method, the data from the underlying database is preprocessed before the analysis by using attribute identification and selection, numerical discretization, and handling of missing values. For solid results, the proposed method needs to be further investigated on diverse diabetes databases. Likewise, in [48], a model is developed for the prediction of hypoglycemia in PH of 30 minutes. In this method, the CGM data is preprocessed and incorporates different decision tree algorithms such as J4.8 and REPtree for classification and prediction. The evaluation results show that J4.8 is better for prediction. Nevertheless, the method is only based on CGM data and does not consider other glucose factors such as insulin, physical activity, etc. Additionally, in this work, only few subjects are studied.

In [36], [49], an SVR based method for subcutaneous glucose concentration prediction in the patients with type 1 diabetes is proposed. The proposed method incorporates CMs to estimate the insulin absorption, gut absorption, and the influences of exercise on insulin dynamics and blood glucose. These estimates along with previous CGM data are then fed to the SVM model for future glucose prediction. The method performs well for short prediction intervals, i.e., 15 and 30 minutes. However, the predictive accuracy decreases with an increase in prediction time. Moreover, the method is evaluated based on only three patients' data, which might not be sufficient for concrete results. Besides, the method does not consider the delay problem in measured value of interstitial glucose and that of plasma glucose. Similarly, in [50], a method based on SVR is proposed for the prediction of hypoglycemia. The method used life events such as exercise, meal intake, and insulin intake along with the CGM data for the glucose levels prediction in 30 and 60 minutes of PH. In this method, different features are extracted from life events and CGM data before the prediction. The evaluation of this method against a baseline method shows improved results. On the other hand, systems for diabetes prediction based on noninvasive measurements are also developed.

In [51], a breath analysis system is designed for the screening of diabetes and prediction of glucose concentrations. The system is composed of chemical sensors, which detects certain biomarkers such as acetone and other volatile organic compounds in the breath and then uses the algorithm based on SVM models to predict blood glucose levels. For screening, it uses SVM with Gaussian kernel, while for prediction; it uses SVR with linear kernel. In this system, the incorporated sensors array is optimized. The system also predicts the HbA1c. In the analysis, the data is preprocessed and features are extracted using PCA. The system is evaluated on real breath samples from both healthy and diabetic subjects. The analysis shows acceptable results. However, the system in the current state is not suitable for practical use.

Furthermore, incorporating the physiological features along with past blood glucose concentrations can lead to better prediction of future glucose concentrations. To this end, in [52], a method is developed for the prediction of future glucose concentrations, which uses past CGM data, insulin intake, and meal intake for prediction. The method incorporates physiological modeling of insulin intake and meal to produce features that are fed to the SVR model for prediction in 30 and 60 minutes of PH. In terms of prediction, the method outperforms the baseline prediction models. However, the method still needs to be evaluated on more robust data for concrete results. Moreover, in [53], ANN and first-order polynomial extrapolation algorithm based short-term glucose prediction method is proposed that aims at better prediction by exploring the meal intake along with the CGM data. The method uses an already published physiological model for the estimation of the effect of carbohydrates intake on the glucose level, which is then used for prediction along with the CGM data. The proposed method is evaluated on 20 simulated and 9 real datasets compared to that of models in [35] and [3], respectively, with improved prediction results. However, in this method, the patient should announce the future meal intake information in advance for the accurate prediction of glucose level. This restriction of advance announcement of meal is resolved in [54], where a jump neural network based prediction scheme is proposed for online short time prediction of future glucose concentrations with a prediction horizon of 30 minutes for type 1 diabetics. In this method, the meal intake information is entered at the time of meal and not in advance. In addition, the inputs of ANN in this method are also fed to the output layer. The proposed scheme

uses past CGM values from the CGM sensors along with the carbohydrates intake data concurrent with meal for the prediction. The scheme is evaluated and compared to an ANN based prediction scheme in [53] with satisfactory forecasting results. However, the scheme lacks generality, since it is trained and optimized for data from a particular CGM sensor with a specific sampling time and not for data from diverse sensors and sample times.

Besides, in [55], an ANN based glucose prediction solution for insulin-dependent diabetics is developed in the form of a mobile software application. Diabetics can use this application on their mobile device for the management of normoglycemia. By using this solution, diabetics can adjust their diet and insulin intake in the perspective of future required insulin predicted by the underlying solution. The future insulin requirement is predicted based on the inputs, such as current glucose level, carbohydrates intake, exercise, and time from the diabetics. The proposed solution also provides the storage of the previous insulin intake, glucose levels, carbohydrates, and exercise data, which could then be sent to a remote endocrinologist for the review of glucose concentration of a particular patient. The solution is evaluated in a trial study in a local hospital with satisfactory feedback from the patients under study. Although the proposed solution is a positive initiative in the field of diabetes management toward real implementation of ANN for prediction in real-time, it is still dependent on user inputs and lacks automation. Table 2 shows the comparison of diabetes diagnosis and prediction techniques based on classification.

B. REGRESSION-BASED TECHNIQUES

Regression is a statistical phenomenon in which a predictive model is built based on the relationship between variables on a given data [63]. Similar to classification-based techniques, various prediction techniques for diabetes have been developed in this domain [3], [64]–[75]. In [3], two algorithms based on first order polynomial and first-order autoregressive (AR) models are developed for the prediction of blood glucose from time series of CGM data. These algorithms are meant for the prediction of hypo/hyperglycemia ahead in time with 30 minutes of PH. The algorithms perform well but with a relatively high prediction delay. Moreover, this work only considers the use of CGM data for forecasting. Similarly, in [64], a data-driven algorithm is implemented for real-time prediction of up-coming glucose variations in 10 and 20 minutes PH respectively. The proposed method is based on an autoregressive model, which also incorporates Kalman filter for filtering the raw CGM data in real-time. The performance analysis of the method shows acceptable results in 10 PH. Like the previous method in [64], this method also considers only the CGM data.

In [76], the authors assess the shortcomings in common statistical modeling approaches used in clinical prediction models. Consequently, they tried to assess other alternatives that could be leveraged instead. Clinical prediction models perform poorly with a good example being a previous model that predicted the chances of getting a mutation in germline DNA mismatch repair genes when colorectal cancer is diagnosed. The model relies on research where 38 mutations were found among 870 participants, while the validation is based on an independent cohort with 35 mutations. The modeling approach used in this case entails selection of predictors out of over 37 candidate predictors in a stepwise paradigm and dichotomization of continuous predictors. On the downside, prediction models tend to depict deficiencies when it is done on the basis of a small number of events and when it is created with common or rather suboptimal statistical techniques. This calls for better modeling approaches that can leverage predictive information available and there is also a need to do research into ways to increase the sample sizes.

In order to reduce the prediction delay, in [65], the feasibility of linear AR models is evaluated by developing three different model scenarios: ordinary AR model without smoothing the data, ordinary AR model with smoothing the data, and regularized AR model with smoothing the data. The study shows that the models with smoothed data and regularized coefficients provide satisfactory results in terms of prediction and prediction delay as compared to the rest of the two scenarios. Moreover, the authors highlight that smoothing of raw data in real-time where one does not have the privilege to smooth the dataset at once is still a challenging problem. Similarly, in [67], the authors attempt to develop and assess universal models based on the data-driven AR model, described in [65], for diabetes prediction in the PH of 30 minutes. Such a model is developed based on the data profile of one diabetic subject and could be used to predict the glucose concentrations in any other subject with any type of diabetes, i.e., type 1 or type 2, regardless of the subject on which the model is originally developed. This is possible due to the invariance property of AR models to the variations in phase and amplitude of the signal (glucose signal) and their only dependency on frequency of the signal. The analysis of the presented study shows the feasibility of the universal models in the area of diabetes management.

Authors in [86] create a metabolomics signature that can predict the likelihood of transitioning from GDM to type 2 diabetics. For this study, a group of 1035 women with GDM pregnancy were placed on a 6-9 weeks' postpartum, then screened for type 2 diabetics yearly for 2 years. It was found that of the 1010 who initially did not have type 2 diabetes at the start of the study, 113 transitioned within two years while another 17 developed type 2 diabetes in 2-4 years. Authors then conducted metabolomics with baseline fasting plasma and recognized 21 metabolites that expressively varied by incident type 2 diabetes status. Using machine learning, authors predicted the development of type 2 diabetes with a discriminative power of 83.0% in the training set and 76.9% in an independent testing set. This is far more effective than using fasting plasma glucose alone. The most recommended procedure after GDM is the type 2 diabetes screening early postpartum via oral glucose tolerance test (OGTT) but tends to be time consuming and

inconvenient unlike the metabolomics signature, which is able to predict type 2 diabetes incidence from a single fasting sample. In this regard, the metabolomics signature can aid in earlier intervention.

Moreover, in [66], an algorithm, namely recursive autoregressive partial least squares (RARPLS), is developed and evaluated for hypoglycemia alarm systems. The algorithm is meant for real systems; thus, it raises an alarm if the current glucose concentration from the CGM system is less than the hypoglycemia threshold, otherwise, it forecasts the future concentration. The prediction performance of the algorithm provides good results compared to the other algorithms in the PH of 30 minutes. However, similar to other techniques, this technique is also developed based on only CGM data. Additionally, since the glucose profile of individuals with diabetes is unstable due to various factors, different simple data-driven AR models could not be able to grasp the trends and volatility in the glucose concentration, and hence fail to provide reliable predictions.

To this end, in [68], generalized autoregressive conditional heteroscedasticity (GARCH) models are explored. Based on the comparative analysis with other higher and lower order AR and state-space models, the authors present that GARCH models are better in grasping the variation in the glucose concentrations and the short time prediction of future glucose levels. Furthermore, in [69], a method for glucose levels prediction based on autoregressive model with exogenous input (ARX) and physiologically inspired adaptive gain is developed. This method incorporates the meal intake by modeling the carbohydrate intake and insulin injection by modeling aspart and detemir plasma insulin for individuals' profiles into the scheme for the forecasting of future glucose concentrations in the 45 minutes of PH. The proposed method shows satisfactory results in the given PH.

On the other hand, in [70], individual subject-specific models are developed for the prediction of future glucose concentration based on diabetics' CGM data, which incorporates linear recursive models such as AR model and autoregressive moving average (ARMA) model. In this work, the models are accompanied with recursive identification and change detection method that enable the models to dynamically adapt to the conditions of intra/inter subject glucose variations and disturbances respectively. The results and analysis show that such inclusions increase the perdition accuracy. This algorithm is also evaluated in [71] in terms of an automated closed-loop insulin system for type 1 diabetes.

Furthermore, in order to develop a global/universal prediction model that could be able to accurately predict the glucose concentrations in inter-subjects without the customization of the model, different modeling attempts are made based on AR linear models in different studies, for example [67], [72]–[74]. In [72], [73], AR models with exogenous inputs, latent variable and glucose dynamics frequency bands are investigated for the online prediction of glucose concentration in type 1 diabetics in order to cope with hypo/hyperglycemia. In the light of this investigation, the objective is to study the development of global AR model for the online prediction of glucose levels and its applicability and feasibility in inter-subjects that show variability in the glucose dynamics. Such a model is developed based on a single subject and then could be applied to other subjects without parameters adjustment. The authors report that the frequency separation based global AR model is feasible. In addition, the authors present that AR and latent variable (LV) with exogenous inputs such as insulin and meal (ARX/LVX) models and with standard subject-dependent model (SD) are more accurate than global ARX/LVX models. However, the global ARX/LVX models are not suitable for the direct use in glucose control. Such a problem of exogenous inputs on glucose variability in inter-subject needs to be further investigated for the development of global model. Similarly, in order to obtain a generalized framework for the modeling of glucose concentrations in any new type 1 diabetics by using a base model, in [74], [75], a method for rapid identification of model for the short-term prediction of online glucose concentrations is presented. In this method, the concept of model migration is used, in which first a base model is empirically developed based on a priori knowledge. Then, models for new subjects are obtained by adjusting the input parameters based on small amount of data from the new subjects in order to capture the inter-subject differences. The method employs ARX for modeling. In addition, the method is evaluated on 30 simulated subjects obtained through UVA/Padova metabolic simulator. The performance evaluation of the proposed method provides comparable results to those of the standard subject-dependent modeling method. However, since the method is evaluated on in-silico subjects, it needs to be further investigated on clinical subjects for concrete results.

Additionally, the work in [77] explored the possibility of using semi-automatically labeled training sets to create phenotype models on the basis of machine learning. The proposed approach uses a patient's medical condition and a list of keywords specific to phenotype to create noisy labeled data. From this, authors then trained L1 penalized logistic regression models that can help detect chronic and acute diseases. Authors then compared the performance of this approach to other state-of-the-art models. For Type 2 diabetes mellitus and myocardial infarction, the proposed approach attained precision and accuracy of 0.90, 0.89, and 0.86, 0.89, respectively.

Similarly, the study in [85] evaluated the possibility of using data mining to establish the desired threshold for glycated hemoglobin (HbA1c) and to assess the possibility of using extra biomarkers coupled with HbA1c to enhance the accuracy of type 2 diabetes mellitus (T2DM) diagnosis. It was found that the accuracy improved when 8-hydroxy-2-deoxyguanosine (8-OhdG) (an oxidative stress marker) and interleukin-6 (IL-6) were included. It was, however, opposite when HbA1c range was between 5.73% and 6.22%. This demonstrates that data analytics approaches together with large clinical datasets can be used to find proper cut off values and the markers then can improve diagnosis of T2DM.

Recent studies on genome association have indicated that the human leukocyte antigen (HLA) genes portray stronger association with a number of autoimmune diseases like type 1 diabetes (T1D). In this regard, authors in [78] built an HLA-based disease predictive model to help reduce this genetic association. The shortcomings of conventional model-building techniques are that they become less optimal especially with highly polymorphic predictors. To overcome this problem, authors put forward an alternative method, which involves taking complex genotypes of HLA genes as objects or exemplars, and then one will concentrate on any association of disease phenotype with the exemplars using similarity measurements. To build a predictive model, authors incorporated both the Kernel representative theorem and the machine learning techniques in a penalized likelihood method used to choose the exemplars associated with the disease. A total of eight HLA genes were taken in building a predictive model for the T1D study. These sampled genes include HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5, HLA-DQA1, HLA-DQB1, HLA-DPA1, and HLA-DPB1. From this predictive model, the values for the area under the curves for the training and validating set were 0.92 and 0.89 respectively.

In [79], data mining techniques were used to create a model for predicting the chances of patients developing liver cancer within 6 years of being diagnosed with type 2 diabetes. The data used in the work was retrieved from the National Health Insurance Research Database (NHIRD) of Taiwan, which has records of over 22M people. From the dataset, the authors were specifically interested in the patients diagnosed with type 2 diabetes between 2000 and 2003, but only those who had not been diagnosed with cancer previously. Authors then did training and testing before creating an ANN and logistic regression (LR) prediction models. Comparing the two models, ANN outperformed LR in a number of performance metrics, namely sensitivity (0.757), specificity (0.755), and the area under the receiver operating characteristic curve (0.873).

Correlation between hypertriglyceridemia waist (HW) phenotype and type 2 diabetes in Korean adults were investigated in [80]. Authors assessed the predictive capabilities of various phenotypes having a blend of individual anthropometric measurements and triglyceride (TG) levels. First, the authors did a measurement of fasting plasma glucose and TG levels. Then, they did anthropometric measurements before using binary LR to find out the statistical difference between subjects with type 2 diabetes and normal ones. The predictive power of the phenotypes was evaluated using two machine learning algorithms, i.e., naive Bayes (NB) and LR. The experiments were done using a 10-fold cross validation approach and it was established that the presence of HW is closely linked with type 2 diabetes. Moreover, the predictive capability of the phenotypes was much more accurate in women than in men.

The work in [81] conducted a series of experiments to determine specific biomarkers that are key to detecting the early stages of diabetic retinopathy (DR), which tends to be quite a common microvascular disorder among diabetes mellitus patients. The first process involved the identification of those biomarker candidates found within the human vitreous. A total of 96 proteins both from the previously published works on DR and their own experimental data were selected. The second step entailed the confirmation and validation of the selected biomarker candidates. Authors extracted plasma from two groups of diabetic patients; those without DR (No DR) and those with mild or moderate non-proliferative diabetic retinopathy (Mi or Mo NPDR). The two samples were first subjected to two reaction processes, namely the semi-quantitative multiple reaction monitoring (SQ-MRM) and stable-isotope dilution multiple reaction monitoring (SID-MRM). In the final validation process, authors conducted a multiplex assay on 15 biomarker candidates identified in the SID-MRM analysis. The results were then compiled and presented in merged AUC values. The No DR versus Mo NPDR scored 0.99, while the No DR versus Mi and Mo NPDR combination yielded 0.93. Authors noted that more samples are needed for accurate results. The comparison of regression-based techniques for diabetes diagnosis and prediction is shown in Table 3.

C. ASSOCIATION-BASED TECHNIQUES

Association rule mining (ARM) is a process of extracting frequent patterns, correlations, and associations between sets of data items in the databases or data repositories [87]. Several association rule based diabetes diagnosis and prediction techniques are presented in the literature [88]–[96]. In [88], ARM is used for the extraction of predictive risk patterns for type 2 diabetes using data from the Tehran lipid and glucose study (TLGS). In order to identify the risk patterns in the data, the study uses 21 different input variables from demographic characteristics, anthropometric measures, and biomedical and disease history of the people in the data, which includes both male and female. This study shows that ARM could be used to determine the occurrence of predictors or variables in people who will develop diabetes in the future. However, the study lacks the inclusion of other factors such as nutritional and sociological factors for the extraction of predictive rules. Similarly, in [92], a tool is developed based on ARM that is used to find association between diabetes (type 2) and its related comorbidities. This system incorporates apriori algorithm for association analysis of clinical diagnosis data. The analysis results of this study show that essential hypertension is the key in the association between type 2 diabetes and its comorbidities. The study uses a single data source and is limited to the association analysis of comorbidities.

On the other hand, in [89], ARM is used on the genetic information of individuals to extract the interaction information of susceptible genes that could provide clue of susceptibility of people to type 2 diabetes. The authors use apriori-gen algorithm to extract the association between the variants of DNA (single nucleotide polymorphisms (SNP))

and type 2 diabetes. Although the study shows fair results, it needs further validation and improvement. Similarly, in [90], a survival association rule mining based method is proposed for the risk assessment of diabetes that incorporates the dosage effects as well as compensates for the confounders, such as age and gender. The performance evaluation of this method provides comparative results to that of other survival models and significant improvement is observed compared to renowned Framingham score. On the contrary, in [91], ARM is used for the discovery of association in diabetes data with ophthalmic complication. The authors use age, gender, and payment method of treatment expense for classification and incorporate apriori algorithm. The purpose of this study is only to explain the association among fundamental parameters in diabetes dataset with complications that could lead to the betterment of healthcare.

Furthermore, in medical records, extensively large risk factors exist, which also produce large sets of association rules that make it difficult to interpret. To this end, different studies are conducted, for example [93], to develop and evaluate rules summarization techniques in order to make more compact and easily interpretable rules. In [93], for the prediction of the risk of diabetes in subpopulations through electronic medical records, four rules summarization techniques (Top-K, APRX-Collection, RPGlobal, and bottom up summarization (BUS)) are analyzed and compared in terms of their applicability, strength, and weaknesses. From the analysis, it is observed that all the techniques provide reasonable results; however, the most accurate technique is the bottom up summarization technique for the risk estimation to subpopulations. Additionally, these techniques are extended by incorporating the information related to continuous outcome variables using survival analysis modeling and distributional association rules.

Table 4 shows the comparison of association-based techniques for diabetes diagnosis and prediction.

D. CLUSTERING-BASED TECHNIQUES

Clustering is an unsupervised learning process in which similar objects are grouped into a cluster, without having any predefined classes as compared to classification [34]. There are abundant clustering techniques for diabetes prediction in the literature, such as [97], [98], [101]–[105], [110]–[115]. In [100], clustering is used to identify the characteristics that could determine the stage and presence of diabetes in a number of women based on the diabetic and normal classified data. After comparing the results of different clustering algorithms, partitioning around medoids (PAM) algorithm is found with best results; hence, incorporated in the study.

It is observed that women with diabetes have unique characteristics than others, which can help in predicting the diabetes stage (initial or advance) and determining the maximum number of women suffering from this disease. The study is generic and employs existing algorithms. Likewise, in [97], a system for diabetes diagnosis and prevention based on hierarchical and conceptual clustering is presented. The system incorporates hierarchy of attributes (life style that leads to the disease, i.e., symptoms) and concepts (type of disease, e.g., type 1, type 2, etc.), and the relationship (disease type and symptoms, disease type and precautions) between the two. Although such kind of system can be useful in diabetes diagnosis and prevention, it needs automation in terms of table generation for the system. Moreover, the system needs attributes, concepts and relational tables, which are not generated automatically. Besides, for large data, the complexity will come into play.

In [106], a novel approach for diagnosing the disease on the basis of a predictive machine learning approach is proposed. The proposed framework is a clustering-based feature extraction framework that uses disease diagnostic datasets. The information is reduced by first establishing disease clusters, a process that makes use of co-occurrence statistics. The clusters are then reduced further in an optimization process, after which the remaining ones are used in a training set. The approach is used to predict the severity of a disease and the chances of readmission. The clustering and feature extraction algorithm was trained on the 2012 National Inpatient Sample (NIS) of Healthcare Cost and Utilization Project (HCUP) dataset that has seven million hospital discharge records and ICD-9-CM codes. In the experimental setup, the algorithm was evaluated on Ronald Reagan UCLA Medical Center Electronic Health Records (EHR) from 3041 Congestive Heart Failure (CHF) patients and the UCI 130-US diabetes dataset. For performance analysis, the algorithm was compared to other popular comorbidity frameworks like the Charlson's index, Elixhauser's comorbidities, and their variations. The results of the evaluation showed that the proposed method had significant gains of 10.7–22.1% in predictive accuracy levels for CHF severity of condition prediction, while it had 4.65–5.75% in diabetes readmission prediction.

In [98], clustering is used to predict the obesity and diabetes rates in a county-level population using demographic, socioeconomic, and environmental variables. This method is suitable for application in a large geographical area for the prediction of diet-related chronic disease such as diabetes in different subpopulations of the area. Similarly, in [99], a divisive hierarchical clustering based algorithm is presented for diabetes identification and prevention. The algorithm clusters patients with similar risk factors that lead to diabetes in order to identify their subpopulation. The performance of the algorithm is evaluated over a large cohort of patients' data with satisfactory results, which outclasses the well-known Framingham score. The results show that the method successfully clusters patients with lower and high risk of diabetes compared to the general population. However, as a prediction tool, the method can be susceptible to the problem of overfitting if not properly handled with the number of patients in each node. Likewise, in [101], a multiple-level clustering based data analysis method is presented for the identification of patients' groups with similar examination history in a dataset with variable data distribution. In this method, the patients' examination history data is represented in a vector space model through term frequency–inverse document frequency

(TF-IDF) method. Moreover, this method uses density-based clustering, which is less sensitive to the outliers and does not require prior knowledge of the number of expected clusters in the underlying data as compared to model-based and partitioning-based clustering. The method is evaluated on a real diabetic data with effective results in terms of patients' groups with similar examination history and increasing severity in the complications of the diabetes disease. However, one drawback could be the high execution time of this method.

Conversely, in [102], [103], a web-enabled grid-based interactive diabetes system (GIDS) is proposed for the analysis and optimized view of diabetes patients' medical data, which is aimed at the prediction of diabetes implications by viewing the medical records of the patients. The proposed system utilizes an agglomerative clustering based algorithm with chronological policy for the correlation analysis of patients' medical data. The system is experimented using physical and virtual computers for the analysis of a single diabetic blood sugar data. The basic idea of the system is to help researchers or medical professionals in analyzing large-scale medical records of diabetic patients. Nevertheless, the system is not evaluated over a large-scale multidimensional data. Besides, the work only explains the workflow of the system and lacks the explanation of how it clusters the corresponding medical records. In [104], [105], a method, similar to the method presented in [102], [103], is proposed for the identification of temporal variations in metabolic syndrome. However, in order to overcome the limitations in [102], [103], such as the identification of value for sensitivity indicator $\acute{\alpha}$ (correlation between observations) and quantitative integration of multiple examination reports, [104], [105] use an areal similarity degree (ASD) based chronological clustering, where the ASD is a similarity-based risk model between two weighted radar charts, which estimates distance between chronologically ordered medical reports. In this method, the distance variance and control range is used to monitor and control the health status of the patients. This study also lacks the evaluation of the method on a large-scale multidimensional data.

In [116], the authors explored the feasibility of using back propagation algorithm in predicting diabetes mellitus. Moreover, the efficacy of other approaches such as J48, Naïve Bayes and SVM were also analyzed in comparison to back propagation. The authors used a 5-fold cross validation technique and a big value learning rate to fine-tune the accuracy levels of the proposed algorithm. In the experimental phase, a [8-6-1] neural network architecture was designed to predict diabetes on the basis of the PIMA Indian dataset. The results showed that the back propagation algorithm had better accuracy levels in predicting diabetes compared to the likes of SVM, J48, and Naïve Bayes algorithm.

Additionally, authors in [108] studied the use of machine learning techniques in detecting the process of fasting blood glucose levels of people. The process involved random sampling of 175 volunteers (50% diabetic and 50% diabetes free). 70% of the data set was used to train the models, while the remaining 30% was used for model testing. The machine learning techniques used in this study include LR, SVM, and ANN. To validate the data points, random shuffling was applied three times for each algorithm before the implementation. Authors chose to represent model performance using four statistical parameters namely accuracy, sensitivity, precision, and F1 score. From the model analysis report, SVM technique using RBF kernel outclassed the other machine learning techniques given that it scored 85% in accuracy, 84% in precision, 85% in sensitivity, and 85% in F1 score.

Finally, an interpretable method for the detection of DR was examined in [109]. Authors added the regression activation map (RAM) to the global averaging pooling layer of the convolutional neural network (CNN) to help obtain the visual-interpretable feature. The importance of RAM in this particular model is to help localize the discriminative regions of a retina image with the aim of displaying the severity levels for the regions of interest. By conducting experiments on a wide scale of the retina image dataset, it was shown that the proposed CNN model achieved better performance of DR detection than the state-of-the-art models. Table 5 shows the comparison of diabetes diagnosis and prediction techniques based on clustering.

E. SEQUENTIAL PATTERN MINING BASED TECHNIQUES

Sequential pattern mining (SPM) is a process of identifying similar patterns [131]. Several SPM based techniques including [117]–[120], [120], [132] have been developed for the diagnoses of diabetes. In [117], an insulin therapy and medical guidelines tool is presented based on mining the clinical data of type 1 diabetes using SPM for the decision support of insulin therapy and the prediction of its impact on blood glucose level. Such a tool can be used by the healthcare professionals to make decisions about the insulin treatment according to the blood glucose level. In this method, the sequences are mined based on template patterns that are pre-defined by the physicians. The algorithm is evaluated on a real clinical data with satisfactory results. However, this method has low sequence generalization capability and uses the patterns separately for each individual patient. Similarly, in [118], differential sequential pattern mining is used to develop a decision support tool for insulin therapy concerning basal insulin in type 1 diabetics with more generalization capabilities as compared to [117]. Besides, [118] also introduces a new way of qualitative evaluation of nocturnal glycemia. However, this work does not consider the bolus insulin therapy.

In [119], a sequential mining technique is used for mining the diabetes data in order to extract knowledge from the data that can be used for managing the disease. In this work, a tree is constructed from the discovered patterns that show the medical events with their frequency and sequence. This study just investigates and explains the use of sequential pattern mining in the diabetes care but lacks the evaluation and concrete results. Moreover, in [120], sequential pattern mining is used for the prediction of temporal relationship

between the medication for diabetics in order to predict the next expected medication to be prescribed for a patient along his/her path of therapy. The authors use CSPADE algorithm for mining the patterns of diabetes prescriptions. Then, based on the relationships that are discovered, rules are defined for the prediction of the next medication. The evaluation of the proposed method shows reasonable results; however, the limitation of this work is the use of Claims Data and the short duration of the dataset. The comparison of sequential pattern mining based techniques for diabetes diagnosis and prediction is shown in Table 6.

F. HYBRID TECHNIQUES

Hybrid prediction techniques are defined as those techniques that incorporate the combination of different models from the same class or different classes. This is the robust class in the area with numerous diagnosis and prediction techniques, for instance [121]–[130], [133], [134]. In [125], a hybrid method for the prediction of type 2 diabetes is presented. In this method, for patterns extraction, K-means clustering algorithm is used, whereas for classification, C4.5 algorithm is used. Similarly, for type 2 diabetes detection, in [121], a classification based association rule mining method is developed using modified particle swarm optimization in combination with least squares SVM (MPSO-LSSVM) and outlier detection method. In the proposed method, frequent item-sets are generated using Complete Frequent Pattern-growth++ (CFP-growth++). Then, the appropriate rules are generated using (MPSO-LSSVM). An outlier detection method is also incorporated for outliers detection in the corresponding generated association rules. The method in [121] is further improved in [122] by using an improved frequent pattern growth (IFP-Growth) with hybrid enhanced artificial bee colony-advanced kernel SVM (HEABC-AKSVM-IFP Growth) classifier for the classification based generation of association rules in the diabetes data. In this system, firstly, the frequent item-sets are generated and then classification based association rules are produced for the prediction origin and prediction of diabetes. The system is evaluated against other classification techniques, such as MPSO-LSSVMCFP Growth++, SVM-FP Growth, and ABC-LSSVM-IFP Growth, in terms of the numbers of derived rules, processing time, and classification accuracy, with improved results.

In the work done in [126], authors looked at a method for identifying GDM at the early phase of pregnancy, thus allowing early response in terms of treatment. It is one of the diseases that has become widespread these days owing to the changing human lifestyle. GDM is chronic and it tends to be spotted at the mid-phase or latter stages of pregnancy, which makes the treatment complicated to both the mother and the child. The work presented by the authors sought to answer the question: ''which one is the best approach for predicting GDM in a timely and accurate fashion?'' In this regard, the authors proposed using a blend of ANN and decision tree to reduce error rates and consequently improve the prediction accuracy and precision levels. The authors also showed that the approach could be used to reduce the mortality rates related to GDM through timely medical intervention. The results of this work also showed the potential intelligent diagnosis systems that could be used in improving the quality of healthcare in the realm of predicting diseases, thus allowing treatment and even prevention.

Furthermore, in [123], a clustering based algorithm is presented for the prediction of susceptibility of individuals to type 2 diabetes. In order to find possible similarity, this cluster-based distance algorithm uses the analysis of genetic information of individuals in comparison to that of other individuals with their disease status already known. For the classification of testing genotype, K nearest neighbors (KNN) scheme is incorporated. The results of the method show good prediction rates on type 2 genotype data. Similarly, in [128], hierarchical clustering support vector machine (HCSVM) model is presented for the classification of type 2 diagnosed patients in a large dataset. In this method, multiple homogeneous clusters are utilized for classification in a given large dataset by partitioning the data into small clusters at multiple levels, which also improves classification accuracy compared to the single SVM model and one-level CSVMs model. However, the overall accuracy of HCSVM is not high enough. Conversely, in [124], a pattern mining based framework is developed and evaluated over a diabetic data with efficient detection of diabetes related disorders. This method is based on recent temporal pattern (RTP), where patterns are mined by starting from the most recent observation patterns to the frequent temporal patterns backward in time. Moreover, instead of mining the frequent RTPs from the entire data with single global minimum support, the frequent RTPs from each class are mined separately using local minimum support. In this method, SVM is used for classification. The evaluation of this framework shows that the method is useful in finding patterns that are vital for the prediction of diabetes related diseases, such as renal infections, cardiological disorders, etc.

Moreover, in [125], a hybrid prediction model is presented for the prediction and screening of type 2 diabetes. The model is based on K-means clustering algorithm and decision tree classification algorithm, where the clustering algorithm extracts the patterns from the original data followed by the classifier with k-fold cross-validation method to classify the newly diagnosed individuals who will be likely to develop the disease in the next five years. The proposed model is evaluated against similar models with improved accuracy. The main contribution of this work is the removal of irrelevant instances from the data for improving the classification accuracy. Similarly, in [127], a hybrid model is used, which incorporates K-means clustering for irrelevant instances removal from the data followed by a genetic algorithm for features selection and SVM classifier for classification using 10-fold cross-validation. The performance evaluation results on the same dataset show that this model achieves higher accuracy compared to the model used in [125].

On the other hand, in [129], a universal adaptive-weighted-average framework is proposed for the prediction of blood glucose concentrations. This framework can combine various algorithms for prediction, which is demonstrated through the combination of AR model, extreme learning machine (ELM), and SVR algorithm. The performance evaluation of the proposed framework against the corresponding individual models shows better results. However, the proposed model only considers the CGM data and is evaluated over a smaller dataset with less data points. Similarly, in [130], an early warning system for hypo/hyperglycemia events in type 1 diabetics is proposed, which incorporates autoregressive model with an output correction module (cARX) and RNN models. The proposed hybrid system shows better prediction and detection accuracy with minimum false alarms for a short prediction horizon compared to the individual models. Table 7 shows the comparison of diabetes diagnosis and prediction techniques based on hybrid class.

V. DISCUSSION AND OPEN ISSUES

This section discusses the various classes of data mining based techniques explained above for diabetes diagnosis and prediction as well as highlights the open issues and challenges for future research in this area.

A. DISCUSSION

Tables 2-7 summarize our analysis and comparison of diabetes diagnosis and prediction techniques by evaluating them on the basis of the various parameters including Algorithm/Model, Data Input, Type of Diabetes, Data Preprocessing, Evaluation Dataset used, Type of Implementation, and Plug-n-Play capability.

Alg/Mod (Algorithm/Model) parameter describes the technique or model used in the proposed schemes. Most of the classification-based schemes use ANN and SVR, etc. Similarly, the association, regression, and clustering-based schemes mostly use data mining based approaches. Moreover, the hybrid schemes mostly evolve by using a combination of different models or schemes, i.e., ANN, data mining etc. Disease classification and detection process sometimes requires multi-disciplinary schemes to achieve high accuracy. The advantage of using hybrid schemes is that these schemes are mostly constituted by using a combination of two different schemes in such a way that the schemes support each other to provide impactful results.

Data Input is the other useful parameter, which describes the input data type that is selected and then passed to the proposed scheme for the prediction and analysis of diabetes. It significantly affects the performance of the prediction, diagnosis and analysis process. Choosing the optimal input data will yield high prediction results. Some of the classification-based schemes involve high human intervention (i.e., manual data inputs in the diary and manual integration of CGM values) during the data input phase; this increases the chances of wrong prediction and diagnosis due to human errors. Similarly, the prediction accuracy is highly affected if the chosen input data does not cover all the key features. We observed that most of the classification-based schemes use a single parameter (i.e., CGM, or insulin, etc.) and hence suffer with low prediction accuracy. Moreover, along with single parameter or factor considerations, some classification-based schemes are strictly dependent on some particular hardware devices; this increases the hurdles for the availability and adaptability of these schemes.

Another excellent option is the type of schemes that use ARM for the prediction and analysis of diabetes. The ARM-based schemes suffer from using very limited or sometimes using only a single parameter or factor for extraction and summarization of the association rules. Moreover, the ARM-based schemes also struggle with validation and maturity. Similarly, auto-regression based schemes are another good choice for diabetes prediction and analysis. Most of these schemes use only CGM data for diabetes prediction and analysis, and hence suffer from low prediction accuracy. This is because the individual's glucose profile has a variable nature depending upon several factors, which makes it very difficult for auto-regression-based schemes to appropriately predict diabetes. Some of the schemes suffer from data smoothing, and hence cannot be used in real-time diabetes prediction and analysis.

In order to classify the diabetic and normal profiles, clustering-based schemes provide very promising results. However, most of the cluster-based schemes struggle with the problems of plug-n-play capability. This means that these schemes have human intervention during the classification and analysis phase. Similarly, some of the cluster-based schemes also face the problems of high execution time, flexibility, and robustness. Sequential pattern mining schemes are mainly used to develop a decision support system to assist physicians and practitioners in the diabetes prediction and analysis.

However, these schemes have pattern generalization issues due to the use of individual patterns for every subject or patient. Moreover, some of the schemes are tested on very short duration datasets; this causes the evaluation to produce insubstantial results. In order to increase the accuracy and performance of the diabetes prediction and analysis, the hybrid type of schemes play an important role. These schemes combine two or more classes of schemes, i.e., clustering and classification, for improved accuracy and performance. Most of these schemes show very promising performance due to the use of accurate functionalities and properties at an optimal point. In this study, we found that most of the hybrid schemes outperform the non-hybrid schemes (all other classes) in terms of accuracy and performance.

B. OPEN ISSUES

The following are some of the future directions we observed during the analysis of the diabetes prediction schemes:

- *Plug-n-play capability:* As discussed, plug-n-play capability makes the scheme free of human errors, as it

TABLE 8. Challenging issues in the field of diabetes diagnosis and prediction.

Field	Area	Challenging Issues
Diabetes data analysis	Data	Availability of relevant accurate and quality data; Data collection and data sharing; Data privacy & security; Data integration from heterogeneous sources; Data access and storage.
	Data cleaning and pre-processing	Appropriate data selection; Data cleaning; Feature selection and extraction; Dimensionality reduction; Data denoising; Data transformation; Data integration
	Diagnosis and prediction techniques	Generic and universal techniques; Clinical and public usability; Evaluation of existing techniques over recent new data sets; Robust software tools; Development of real-time online detection and prediction mining tools; Selection of appropriate models; Integration of models from different domains; Efficiency and accuracy

reduces the human involvement in the prediction and analysis process. Almost all of the schemes suffer from plug-n-play capability, as these schemes involve human intervention at a certain stage in the analysis and prediction process. In order to increase the applicability and adaptability of the diabetes prediction and analysis system, an optimal plug-n-play capability is required. In a plug-n-play system, a user simply needs to switch the system on and it starts its functionality, which makes the system easy to use and operate. The plug-n-play capability increases the usability and scalability of the scheme [135].

- *Multi-parameter input data:* Accurate diabetes prediction and analysis depends on several key factors. Combining these factors and properties for passing to the diabetes prediction system will increase the chances of accurate and precise prediction. It is evident from the analysis of the schemes in all classes that most of them suffer from either single data input parameter or the parameter selection process is not optimal. In any case, the scheme will not perform precisely in terms of prediction accuracy.

Table 8 shows the various challenging issues related to the data, data cleaning, and pre-processing, as well as the techniques used for diagnosis and prediction of diabetes.

VI. CONCLUSION

In this paper, we presented a comprehensive review of the state-of-the-art on the glycemic control in the domains of data mining based diabetes diagnosis and prediction techniques and their classification based on the underlying models used. Based on the literature review of data mining based techniques for diabetes detection, classification and prediction, we provide a comprehensive classification of the commonly used diabetes diagnosis and prediction techniques. Moreover, we evaluated different schemes on parameters like, algorithm/model, type of input data (data input), plug-n-play capability, etc. On the basis of this analysis and evaluation, we conclude that for accurate detection, classification, and prediction of the disease, we need to preprocess the data and use hybrid techniques, which incorporate different models in parallel instead of using an individual model. For preprocessing, we need to use dimensionality reduction, denoising, feature selection, and feature extraction techniques in combination with the classification and prediction schemes for optimal performance and results.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research and Innovation, ''Ministry of Education'' in Saudi Arabia for funding this research work through the Project no. (IFKSURP-255).

REFERENCES

[1] P. Dua, F. J. Doyle, and E. N. Pistikopoulos, ''Model-based blood glucose control for type 1 diabetes via parametric programming,'' *IEEE Trans. Biomed. Eng.*, vol. 53, no. 8, pp. 1478–1491, Aug. 2006.

[2] American Diabetes Association, ''2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2020,'' *Diabetes Care*, vol. 43, no. 1, pp. S14–S31, Jan. 2020.

[3] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, ''Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series,'' *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 931–937, May 2007.

[4] S. Guerra, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, ''Enhancing the accuracy of subcutaneous glucose sensors: A real-time deconvolution-based approach,'' *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1658–1669, Jun. 2012.

[5] J. M. Norris, R. K. Johnson, and L. C. Stene, ''Type 1 diabetes—Early life origins and changing epidemiology,'' *Lancet Diabetes Endocrinol.*, vol. 8, no. 3, pp. 226–238, Mar. 2020.

[6] *National Diabetes Statistics Report, 2020.* Accessed: Jan. 15, 2021. [Online]. Available: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[7] ID Federation. *IDF DIABETES ATLAS 9th Edition 2019.* Accessed: Jan. 15, 2021. [Online]. Available: https://diabetesatlas.org/en/

[8] L. Olansky and L. Kennedy, ''Finger-stick glucose monitoring: Issues of accuracy and specificity,'' *Diabetes Care*, vol. 33, no. 4, pp. 948–949, Apr. 2010.

[9] J. B. Buse, D. J. Wexler, A. Tsapas, P. Rossing, G. Mingrone, C. Mathieu, D. A. D'Alessio, and M. J. Davies, ''2019 update to: Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD),'' *Diabetologia*, vol. 63, no. 2, pp. 221–228, Feb. 2020.

[10] M. Langendam, Y. M. Luijf, L. Hooft, J. H. D. Vries, A. H. Mudde, and R. J. Scholten, ''Continuous glucose monitoring systems for type 1 diabetes mellitus,'' *Cochrane Database Syst. Rev.*, vol. 2012, no. 1, pp. 1–144, 2012, Art. no. CD008101.

[11] C. Choleau, J. C. Klein, G. Reach, B. Aussedat, V. Demaria-Pesce, G. S. Wilson, R. Gifford, and W. K. Ward, ''Calibration of a subcutaneous amperometric glucose sensor: Part 1. Effect of measurement uncertainties on the determination of sensor sensitivity and background current,'' *Biosensors Bioelectronics*, vol. 17, no. 8, pp. 641–646, Aug. 2002.

[12] D. Sisodia and D. S. Sisodia, ''Prediction of diabetes using classification algorithms,'' *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, Jun. 2018.

[13] H. Kaur and V. Kumar, ''Predictive modelling and analytics for diabetes using a machine learning approach,'' *Appl. Comput. Inform.*, vol. 16, pp. 1–11, Jul. 2020.

[14] K. Kincade, ''Data mining: Digging for healthcare gold,'' *Insurance Technol.*, vol. 23, no. 2, no. 2, pp. 2–7, 1998.

[15] B. L. Shivakumar and S. Alby, "A survey on data-mining technologies for prediction and diagnosis of diabetes," in *Proc. Int. Conf. Intell. Comput. Appl.*, Mar. 2014, pp. 167–173.
[16] K. Rajalakshmi and D. S. S. Dhenakaran, "Analysis of data mining prediction techniques in healthcare management system," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 5, no. 4, pp. 1343–1347, Apr. 2015.
[17] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, "Data-mining technologies for diabetes: A systematic review," *J. Diabetes Sci. Technol.*, vol. 5, no. 6, pp. 1549–1556, Nov. 2011.
[18] M. Durairaj and K. Priya, "Breast cancer prediction using soft computing techniques a survey," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 8, pp. 135–145, Aug. 2018.
[19] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Comput. Sci.*, vol. 47, pp. 45–51, May 2015.
[20] T. Pala and A. Y. Camurcu, "Evaluation of data mining classification and clustering techniques for diabetes," *Malaysian J. Comput.*, vol. 2, no. 1, pp. 1–9, 2014.
[21] R. M. Rahman and F. Afroz, "Comparison of various classification techniques using different data mining tools for diabetes diagnosis," *J. Softw. Eng. Appl.*, vol. 6, no. 3, p. 85, 2013.
[22] V. Karthikeyani, I. P. Begum, K. Tajudin, and I. S. Begam, "Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction," *Int. J. Comput. Appl.*, vol. 60, no. 12, pp. 26–31, Dec. 2012.
[23] P. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus: A case study," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 1, pp. 8–13, Jan. 2015.
[24] G. Visalatchi, S. J. Gnanasoundhari, and M. Balamurugan, "A survey on data mining methods and techniques for diabetes mellitus," *Int. J. Comput. Sci. Mobile Appl.*, vol. 2, no. 2, pp. 100–105, 2014.
[25] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017.
[26] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *Int. J. Bio-Sci. Bio-Technol.*, vol. 5, no. 5, pp. 241–266, Oct. 2013.
[27] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy Buildings*, vol. 49, pp. 560–567, Jun. 2012.
[28] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, "Real-data comparison of data mining methods in prediction of diabetes in Iran," *Healthcare Informat. Res.*, vol. 19, no. 3, no. 3, pp. 177–185, 2013.
[29] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, "Application of classification based data mining technique in diabetes care," *J. Appl. Sci.*, vol. 13, no. 3, pp. 416–422, Jan. 2013.
[30] E. I. Georga, D. I. Fotiadis, and V. C. Protopappas, "Glucose prediction in type 1 and type 2 diabetic patients using data driven techniques," in *Knowledge-Oriented Applications in Data Mining*. London, U.K.: IntechOpen, Jan. 2011, pp. 277–296.
[31] D. Tomar and S. Agarwal, "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes," *Adv. Artif. Neural Syst.*, vol. 2015, pp. 1–10, Jan. 2015.
[32] D. D. A. Kumar and R. Govindasamy, "Performance and evaluation of classification data mining techniques in diabetes," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1312–1319, 2015.
[33] R. Sheikhpour and M. A. Sarram, "Diagnosis of diabetes using an intelligent approach based on bi-level dimensionality reduction and classification algorithms," *Iranian J. Diabetes Obesity*, vol. 6, no. 2, pp. 74–84, 2014.
[34] J. Han, M. Kamber, and J. Pei, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.
[35] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. J. Gómez, M. Rigla, A. D. Leiva, and M. E. Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes Technol. Therapeutics*, vol. 12, no. 1, pp. 81–88, Jan. 2010.
[36] E. I. Georga, V. C. Protopappas, and D. Polyzos, "Prediction of glucose concentration in type 1 diabetic patients using support vector regression," in *Proc. 10th IEEE Int. Conf. Inf. Technol. Appl. Biomed.*, Nov. 2010, pp. 1–4.
[37] S. G. Mougiakakou, A. Prountzou, D. Iliopoulou, K. S. Nikita, A. Vazeou, and C. S. Bartsocas, "Neural network based glucose–insulin metabolism models for children with type 1 diabetes," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2006, pp. 3545–3548.
[38] S. M. Pappada, M. J. Borst, B. D. Cameron, R. E. Bourey, J. D. Lather, D. Shipp, A. Chiricolo, and T. J. Papadimos, "Development of a neural network model for predicting glucose levels in a surgical critical care setting," *Patient Saf. Surg.*, vol. 4, no. 1, p. 15, 2010.
[39] S. M. Pappada, B. D. Cameron, and P. M. Rosman, "Development of a neural network for prediction of glucose concentration in type 1 diabetes patients," *J. Diabetes Sci. Technol.*, vol. 2, no. 5, pp. 792–801, Sep. 2008.
[40] S. M. Pappada, B. D. Cameron, P. M. Rosman, R. E. Bourey, T. J. Papadimos, W. Olorunto, and M. J. Borst, "Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes," *Diabetes Technol. Therapeutics*, vol. 13, no. 2, pp. 135–141, Feb. 2011.
[41] F. Allam, Z. Nossair, H. Gomma, I. Ibrahim, and M. A.-E. Salam, "Prediction of subcutaneous glucose concentration for type-1 diabetic patients using a feed forward neural network," in *Proc. Int. Conf. Comput. Eng. Syst.*, Nov. 2011, pp. 129–133.
[42] Z. Zainuddin, O. Pauline, and C. Ardil, "A neural network approach in predicting the blood glucose level for diabetic patients," *Int. J. Comput. Intell*, vol. 5, no. 1, pp. 72–79, 2009.
[43] G. Robertson, E. D. Lehmann, W. Sandham, and D. Hamilton, "Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: A proof-of-concept pilot study," *J. Electr. Comput. Eng.*, vol. 2011, pp. 1–11, May 2011.
[44] M. Wiley, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, "Automatic detection of excessive glycemic variability for diabetes management," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, Dec. 2011, pp. 148–154.
[45] F. Ståhl, R. Johansson, and E. Renard, "Ensemble glucose prediction in insulin-dependent diabetes," in *Data-driven Modeling for Diabetes*. Berlin, Germany: Springer, 2014, pp. 37–71.
[46] X. Mo, Y. Wang, and X. Wu, "Hypoglycemia prediction using extreme learning machine (ELM) and regularized ELM," in *Proc. 25th Chin. Control Decis. Conf. (CCDC)*, May 2013, pp. 4405–4409.
[47] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *Proc. Int. Conf. Innov. Inf. Technol.*, Apr. 2011, pp. 303–307.
[48] K. S. Eljil, G. Qadah, and M. Pasquier, "Predicting hypoglycemia in diabetic patients using data mining techniques," in *Proc. 9th Int. Conf. Innov. Inf. Technol. (IIT)*, Mar. 2013, pp. 130–135.
[49] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "Predictive modeling of glucose metabolism using free-living data of type 1 diabetic patients," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 589–592.
[50] C. Marling, M. Wiley, R. Bunescu, J. Shubrook, and F. Schwartz, "Emerging applications for intelligent diabetes management," *AI Mag.*, vol. 33, no. 2, p. 67, Mar. 2012.
[51] K. Yan, D. Zhang, D. Wu, H. Wei, and G. Lu, "Design of a breath analysis system for diabetes screening and blood glucose level prediction," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 11, pp. 2787–2795, Nov. 2014.
[52] R. Bunescu, N. Struble, C. Marling, J. Shubrook, and F. Schwartz, "Blood glucose level prediction using physiological models and support vector regression," in *Proc. 12th Int. Conf. Mach. Learn. Appl.*, vol. 1, Dec. 2013, pp. 135–140.
[53] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1550–1560, Jun. 2012.
[54] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, "Jump neural network for online short-time prediction of blood glucose from continuous monitoring sensors and meal information," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 144–152, Jan. 2014.
[55] K. Curran, E. Nichols, E. Xie, and R. Harper, "An intensive insulinotherapy mobile phone application built on artificial intelligence techniques," *J. Diabetes Sci. Technol.*, vol. 4, no. 1, pp. 209–220, Jan. 2010.
[56] N. K. Kumar, D. Vigneswari, M. V. Krishna, and G. P. Reddy, "An optimized random forest classifier for diabetes mellitus," in *Emerging Technologies in Data Mining and Information Security*. Singapore: Springer, 2019, pp. 765–773.
[57] A. S. Martinez-Vernon, J. A. Covington, R. P. Arasaradnam, S. Esfahani, N. O'Connell, I. Kyrou, and R. S. Savage, "An improved machine learning pipeline for urinary volatiles disease detection: Diagnosing diabetes," *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0204425.
[58] H. Das, B. Naik, and H. Behera, "Classification of diabetes mellitus disease (DMD): A data mining (DM) approach," in *Progress in Computing, Analytics and Networking*. Singapore: Springer, 2018, pp. 539–549.

[59] E. V. Carrera, A. Gonzalez, and R. Carrera, "Automated detection of diabetic retinopathy using SVM," in *Proc. IEEE 24 Int. Conf. Electron., Electr. Eng. Comput. (INTERCON)*, Aug. 2017, pp. 1–4.

[60] S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Informat.*, vol. 59, pp. 185–200, Feb. 2016.

[61] A. Ali and F. A. Khan, "Key agreement schemes in wireless body area networks: Taxonomy and state-of-the-art," *J. Med. Syst.*, vol. 39, no. 10, pp. 1–14, Oct. 2015.

[62] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informat. Med. Unlocked*, vol. 10, pp. 100–107, Jan. 2018.

[63] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. Cambridge, MA, USA: MIT Press, 2001.

[64] Y. Lu, S. Rajaraman, W. K. Ward, R. A. Vigersky, and J. Reifman, "Predicting human subcutaneous glucose concentration in real time: A universal data-driven approach," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 7945–7948.

[65] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting subcutaneous glucose concentration in humans: Data-driven glucose modeling," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 246–254, Feb. 2009.

[66] E. S. Bayrak, K. Turksoy, A. Cinar, L. Quinn, E. Littlejohn, and D. Rollins, "Hypoglycemia early alarm systems based on recursive autoregressive partial least squares models," *J. Diabetes Sci. Technol.*, vol. 7, no. 1, pp. 206–214, Jan. 2013.

[67] A. Gani, A. V. Gribok, Y. Lu, W. K. Ward, R. A. Vigersky, and J. Reifman, "Universal glucose models for predicting subcutaneous glucose concentration in humans," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 157–165, Jan. 2010.

[68] S. K. Paul and M. Samanta, "Predicting upcoming glucose levels in patients with type 1 diabetes using a generalized autoregressive conditional heteroscedasticity modelling approach," *Int. J. Statist. Med. Res.*, vol. 4, no. 2, p. 188, 2015.

[69] G. C. Estrada, H. Kirchsteiger, L. D. Re, and E. Renard, "Innovative approach for online prediction of blood glucose profile in type 1 diabetes patients," in *Proc. Amer. Control Conf.*, Jun. 2010, pp. 2015–2020.

[70] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith, "Estimation of future glucose concentrations with subject-specific recursive linear models," *Diabetes Technol. Therapeutics*, vol. 11, no. 4, pp. 243–253, Apr. 2009.

[71] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith, "Adaptive control strategy for regulation of blood glucose levels in patients with type 1 diabetes," *J. Process Control*, vol. 19, no. 8, pp. 1333–1346, Sep. 2009.

[72] C. Zhao, E. Dassau, H. C. Zisser, L. Jovanovic, F. J. Doyle, and D. E. Seborg, "Online prediction of subcutaneous glucose concentration for type 1 diabetes using empirical models and frequency-band separation," *AIChE J.*, vol. 60, no. 2, pp. 574–584, Feb. 2014.

[73] C. Zhao, Y. Sun, and L. Zhao, "Interindividual glucose dynamics in different frequency bands for online prediction of subcutaneous glucose concentration in type 1 diabetic subjects," *AIChE J.*, vol. 59, no. 11, pp. 4228–4240, Nov. 2013.

[74] C. Zhao and C. Yu, "Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type I diabetes," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1333–1344, May 2015.

[75] C. Yu and C. Zhao, "Rapid model identification for online glucose prediction of new subjects with type 1 diabetes using model migration method," in *Proc. 19th World Congr. Int. Fed. Autom. Control*, Aug. 2014, vol. 19, no. 1, pp. 2094–2099.

[76] E. W. Steyerberg, H. Uno, J. P. A. Ioannidis, B. van Calster, C. Ukaegbu, T. Dhingra, S. Syngal, and F. Kastrinos, "Poor performance of clinical prediction models: The harm of commonly applied methods," *J. Clin. Epidemiol.*, vol. 98, pp. 133–143, Jun. 2018.

[77] V. Agarwal, T. Podchiyska, J. M. Banda, V. Goel, T. I. Leung, E. P. Minty, T. E. Sweeney, E. Gyang, and N. H. Shah, "Learning statistical models of phenotypes using noisy labeled training data," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 6, pp. 1166–1173, Nov. 2016.

[78] L. P. Zhao, H. Bolouri, M. Zhao, D. E. Geraghty, Å. Lernmark, and Better Diabetes Diagnosis Study Group, "An object-oriented regression for building disease predictive models with multiallelic HLA genes," *Genet. Epidemiol.*, vol. 40, no. 4, pp. 315–332, May 2016.

[79] H.-H. Rau, C.-Y. Hsu, Y.-A. Lin, S. Atique, A. Fuad, L.-M. Wei, and M.-H. Hsu, "Development of a Web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network," *Comput. Methods Programs Biomed.*, vol. 125, pp. 58–65, Mar. 2016.

[80] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 39–46, Jan. 2016.

[81] J. Jin, H. Min, S. J. Kim, S. Oh, K. Kim, H. G. Yu, T. Park, and Y. Kim, "Development of diagnostic biomarkers for detecting diabetic retinopathy at early stages using quantitative proteomics," *J. Diabetes Res.*, vol. 2016, pp. 1–22, Nov. 2016.

[82] P. C. Austin, D. van Klaveren, Y. Vergouwe, D. Nieboer, D. S. Lee, and E. W. Steyerberg, "Geographic and temporal validity of prediction models: Different approaches were useful to examine model performance," *J. Clin. Epidemiol.*, vol. 79, pp. 76–85, Nov. 2016.

[83] F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili, "A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results," *J. Clin. Epidemiol.*, vol. 71, pp. 76–85, Mar. 2016.

[84] W. Oh, E. Kim, M. R. Castro, P. J. Caraballo, V. Kumar, M. S. Steinbach, and G. J. Simon, "Type 2 diabetes mellitus trajectories and associated risks," *Big Data*, vol. 4, no. 1, pp. 25–30, Mar. 2016.

[85] H. F. Jelinek, A. Stranieri, A. Yatsko, and S. Venkatraman, "Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis," *Comput. Biol. Med.*, vol. 75, pp. 90–97, Aug. 2016.

[86] A. Allalou, A. Nalla, K. J. Prentice, Y. Liu, M. Zhang, F. F. Dai, X. Ning, L. R. Osborne, B. J. Cox, E. P. Gunderson, and M. B. Wheeler, "A predictive metabolic signature for the transition from gestational diabetes mellitus to type 2 diabetes," *Diabetes*, vol. 65, no. 9, pp. 2529–2539, Sep. 2016.

[87] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993.

[88] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh, "An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database," *Int. J. Endocrinol. Metabolism*, vol. 13, no. 2, pp. 1–9, Apr. 2015.

[89] W. Mao and J. Mao, "The application of apriori-gen algorithm in the association study in type 2 diabetes," in *Proc. 3rd Int. Conf. Bioinf. Biomed. Eng.*, Jun. 2009, pp. 1–4.

[90] G. J. Simon, J. Schrom, M. R. Castro, P. W. Li, and P. J. Caraballo, "Survival association rule mining towards type 2 diabetes risk assessment," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2013, p. 1293.

[91] P. Kasemthaweesab and W. Kurutach, "Study of diabetes mellitus (DM) with ophthalmic complication using association rules of data mining technique," in *Computational Collective Intelligence. Technologies and Applications*. Berlin, Germany: Springer, 2011, pp. 527–536.

[92] H. S. Kim, A. M. Shin, M. K. Kim, and Y. N. Kim, "Comorbidity study on type 2 diabetes mellitus using data mining," *Korean J Intern Med*, vol. 27, no. 2, pp. 197–202, 2012.

[93] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending association rule summarization techniques to assess risk of diabetes mellitus," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 130–141, Jan. 2015.

[94] C.-L. Chan, C.-W. Chen, and B.-J. Liu, "Discovery of association rules in metabolic syndrome related diseases," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 856–862.

[95] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Association rule for classification of type-2 diabetic patients," in *Proc. 2nd Int. Conf. Mach. Learn. Comput.*, 2010, pp. 330–334.

[96] H. Zheng, H. W. Park, and K. H. Ryu, "An efficient association rule mining method to predict diabetes mellitus: KNHANES 2013-2015," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*. Cham, Switzerland: Springer, 2020, pp. 241–249.

[97] S. C. Suh and G. P. Vudumula, "The role of conceptual hierarchies in the diagnosis and prevention of diabetes," in *Proc. 7th Int. Conf. Netw. Comput. Adv. Inf. Manage. (NCM)*, Jun. 2011, pp. 267–275.

[98] A. Flynt and M. I. G. Daepp, "Diet-related chronic disease in the northeastern united states: A model-based clustering approach," *Int. J. Health Geographics*, vol. 14, no. 1, pp. 1–14, Dec. 2015.

[99] E. Kim, W. Oh, D. S. Pieczkiewicz, M. R. Castro, P. J. Caraballo, and G. J. Simon, "Divisive hierarchical clustering towards identifying clinically significant pre-diabetes subpopulations," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2014, p. 1815.

[100] P. Padmaja, S. Vikkurty, N. I. Siddiqui, P. Dasari, B. Ambica, V. V. Rao, M. V. Shaik, and V. R. Rudraraju, "Characteristic evaluation of diabetes data using clustering techniques," *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 11, pp. 244–251, 2008.
[101] D. Antonelli, E. Baralis, G. Bruno, T. Cerquitelli, S. Chiusano, and N. Mahoto, "Analysis of diabetic patients through their examination history," *Expert Syst. Appl.*, vol. 40, no. 11, pp. 4672–4678, Sep. 2013.
[102] F. A. Hazemi, C. H. Youn, and K. A. Al-Rubeaan, "Grid-based interactive diabetes system," in *Proc. IEEE 1st Int. Conf. Healthcare Informat., Imag. Syst. Biol.*, Jul. 2011, pp. 258–263.
[103] F. Al-Hazemi, "Grid-based workflow system for chronic disease study," *Life Sci. J.*, vol. 11, no. 7, pp. 1–3, 2014.
[104] S. Jeong, C.-H. Youn, and Y.-W. Kim, "A method for identifying temporal progress of chronic disease using chronological clustering," in *Proc. IEEE 15th Int. Conf. e-Health Netw., Appl. Services (Healthcom)*, Oct. 2013, pp. 329–333.
[105] S. Jeong, C.-H. Youn, Y.-W. Kim, and S.-O. Shim, "Temporal progress model of metabolic syndrome for clinical decision support system," *IRBM*, vol. 35, no. 6, pp. 310–320, Dec. 2014.
[106] C. Sideris, M. Pourhomayoun, H. Kalantarian, and M. Sarrafzadeh, "A flexible data-driven comorbidity feature extraction framework," *Comput. Biol. Med.*, vol. 73, pp. 165–172, Jun. 2016.
[107] R. Hoyt, S. Linnville, S. Thaler, and J. Moore, "Digital family history data mining with neural networks: A pilot study," *Perspectives Health Inf. Manage.*, vol. 13, no. 1c, pp. 1–14, Jan. 2016.
[108] S. Malik, R. Khadgawat, S. Anand, and S. Gupta, "Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva," *SpringerPlus*, vol. 5, no. 1, p. 701, Dec. 2016.
[109] Z. Wang and J. Yang, "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation," 2017, *arXiv:1703.10757*. [Online]. Available: http://arxiv.org/abs/1703.10757
[110] V. Lakshmi and K. Thilagavathi, "An approach for prediction of diabetic disease by using b-colouring technique in clustering analysis," *Int. J. Appl. Math. Res.*, vol. 1, no. 4, pp. 520–530, Aug. 2012.
[111] S. Vijayarani and M. P. Jothi, "Hierarchical and partitioning clustering algorithms for detecting outliers in data streams," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 3, no. 4, pp. 1–5, Apr. 2014.
[112] R. Awasthi, A. Kumar, and S. Pathak, "An analysis of density based clustering technique with dimensionality reduction for diabetic patient," *Int. J. Comput. Eng. Appl.*, vol. 9, no. 4, pp. 165–171, Apr. 2015.
[113] G. G. Rajput and P. N. Patil, "Detection and classification of exudates using K-means clustering in color retinal images," in *Proc. 5th Int. Conf. Signal Image Process.*, Jan. 2014, pp. 126–130.
[114] R. Paul and A. S. M. L. Hoque, "Clustering medical data to predict the likelihood of diseases," in *Proc. 5th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Jul. 2010, pp. 44–49.
[115] M. S. Kadhm, I. W. Ghindawi, and D. E. Mhawi, "An accurate diabetes prediction system based on K-means clustering and proposed classification approach," *Int. J. Appl. Eng. Res.*, vol. 13, no. 6, pp. 4038–4041, 2018.
[116] F. G. Woldemichael and S. Menaria, "Prediction of diabetes using data mining techniques," in *Proc. 2nd Int. Conf. Trends Electron. Informat. (ICOEI)*, May 2018, pp. 414–418.
[117] W. Froelich, R. Deja, and G. Deja, "Mining therapeutic patterns from clinical data for juvenile diabetes," *Fundamenta Informaticae*, vol. 127, nos. 1–4, pp. 513–528, 2013.
[118] R. Deja, W. Froelich, and G. Deja, "Differential sequential patterns supporting insulin therapy of new-onset type 1 diabetes," *Biomed. Eng. OnLine*, vol. 14, no. 1, p. 13, Dec. 2015.
[119] S. B. Rahaman and M. Shashi, "Sequential mining equips e-health with knowledge for managing diabetes," in *Proc. 4th Int. Conf. New Trends Inf. Sci. Service Sci. (NISS)*, May 2010, pp. 65–71.
[120] A. P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig, "The use of sequential pattern mining to predict next prescribed medications," *J. Biomed. Informat.*, vol. 53, pp. 73–80, Feb. 2015.
[121] T. Karthikeyan and K. Vembandasamy, "A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method," *Indian J. Sci. Technol.*, vol. 8, no. 8, pp. 310–320, 2015.
[122] T. Karthikeyan, K. Vembandasamy, and B. Raghavan, "An intelligent type-II diabetes mellitus diagnosis approach using improved FP-growth with hybrid classifier based arm," *Res. J. Appl. Sci., Eng. Technol.*, vol. 11, no. 5, pp. 549–558, Oct. 2015.
[123] H. Hu and W. Mao, "The application of cluster analysis in type II diabetes genome association study," *J. Comput. Commun.*, vol. 2, no. 9, pp. 1–8, 2014.
[124] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, "Mining recent temporal patterns for event detection in multivariate time series data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 280–288.
[125] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for type-2 diabetic patients," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8102–8108, Dec. 2010.
[126] S. Rouhani and M. MirSharif, "Data mining approach for the early risk assessment of gestational diabetes mellitus," *Int. J. Knowl. Discovery Bioinf.*, vol. 8, no. 1, pp. 1–11, Jan. 2018.
[127] T. Santhanam and M. S. Padmavathi, "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," *Procedia Comput. Sci.*, vol. 47, pp. 76–83, May 2015.
[128] W. Zhong, R. Chow, R. Stolz, J. He, and M. Dowell, "Hierarchical clustering support vector machines for classifying type-2 diabetes patients," in *Bioinformatics Research and Applications*. Springer, 2008, pp. 379–389.
[129] Y. Wang, X. Wu, and X. Mo, "A novel adaptive-weighted-average framework for blood glucose prediction," *Diabetes Technol. Therapeutics*, vol. 15, no. 10, pp. 792–801, Oct. 2013.
[130] E. Daskalaki, S. Mougiakakou, K. Nørgaard, T. Züger, A. Prountzou, and P. Diem, "An early warning system for hypoglycemic/hyperglycemic events based on fusion of adaptive prediction models," *J. Diabetes Sci. Technol.*, vol. 7, no. 3, pp. 689–698, May 2013.
[131] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. Data Eng.*, Jun. 1995, pp. 3–14.
[132] G.-C. Lan, C.-H. Lee, Y.-Y. Lee, V. S. Tseng, J.-S. Wu, C.-Y. Chin, M.-L. Day, S.-C. Wang, C.-N. Chang, and S.-Y. Cheng, "Disease risk prediction by mining personalized health trend patterns: A case study on diabetes," in *Proc. Conf. Technol. Appl. Artif. Intell.*, Nov. 2012, pp. 27–32.
[133] J. N. Mamman, M. B. Abdullahi, A. M. Aibinu, and I. M. Abdullahi, "Diabetes classification using cascaded data mining technique," *Int. J. Comput. Trends Technol.*, vol. 22, no. 2, pp. 53–63, Apr. 2015.
[134] R. Sanakal and S. T. Jayakumari, "Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine," *Int. J. Comput. Trends Technol.*, vol. 11, no. 2, pp. 94–98, 2014.
[135] A. Ali and F. A. Khan, "A broadcast-based key agreement scheme using set reconciliation for wireless body area networks," *J. Med. Syst.*, vol. 38, no. 5, pp. 1–12, May 2014.

FARRUKH ASLAM KHAN (Senior Member, IEEE) received the M.S. degree in computer system engineering from the GIK Institute of Engineering Sciences and Technology, Pakistan, in 2003, and the Ph.D. degree in computer engineering from Jeju National University, South Korea, in 2007. He is currently working as a Professor with the Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia. He also received professional trainings from the Massachusetts Institute of Technology, New York University, IBM, and other institutions. He was the Founding Director of the Wireless Networking and Security (WiNGS) Research Group, National University of Computer and Emerging Sciences, Islamabad, Pakistan. He has published more than 110 research articles in refereed international journals and conferences. He has supervised/co-supervised five Ph.D. students and seventeen M.S. thesis students. Several M.S. and Ph.D. students are also currently working under his supervision. His research interests include cybersecurity, wireless sensor networks and e-Health, bio-inspired and evolutionary computation, and the Internet of Things. He is on the panel of reviewers of over 40 reputed international journals and numerous international conferences. He has co-organized several international conferences and workshops. He is also a Fellow of the British Computer Society (BCS). He serves/served as an Associate Editor for prestigious international journals, including IEEE Access, *PLOS One*, *Neurocomputing* (Elsevier), *Ad Hoc and Sensor Wireless Networks*, *KSII Transactions on Internet and Information Systems*, *Human-Centric Computing and Information Sciences* (Springer), and *Complex & Intelligent Systems* (Springer).

KHAN ZEB received the B.Sc. degree in telecommunication engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2010, and the M.Sc. degree in electrical engineering from King Saud University, Riyadh, Saudi Arabia, in 2015. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Concordia University, Montreal, QC, Canada. From 2010 to 2011, he was a Transmission Network Engineer with LCC (Pvt.) Ltd., Islamabad, Pakistan. From 2011 to 2014, he was a Research Assistant with the Department of Electrical Engineering, College of Engineering, King Saud University. From 2015 to 2016, he was a Researcher with the Center of Excellence in Information Assurance (CoEIA), King Saud University. He has been a Ph.D. Student with the Advanced Electronics and Photonics (AEP) Research Centre, National Research Council (NRC), Ottawa, ON, Canada, since 2018. His research interests include next generation fiber-wireless integrated systems/networks, 5G/6G fronthaul transmission systems, radio-over-fiber technology, microwave photonics, photonic millimeter-wave, space division multiplexing optical transmission technology, quantum-dot (QD) semiconductor lasers and their applications, cyber security, network traffic analysis, and anomaly detection.

MABROOK AL-RAKHAMI (Member, IEEE) received the master's degree in information systems from King Saud University, Riyadh, Saudi Arabia, where he is currently pursuing the Ph.D. degree with the Information Systems Department, College of Computer and Information Sciences. He worked as a Lecturer and taught many courses, such as programming languages in the College of Computer and Information Sciences, King Saud University, Muzahimiyah Branch. He has authored several articles in peer-reviewed journals (IEEE/ACM/Springer/Wiley) and conferences. His research interests include edge intelligence, social networks, cloud computing, the Internet of Things, big data, and health informatics.

ABDELOUAHID DERHAB received the engineering, master's, and Ph.D. degrees in computer science from the University of Sciences and Technology Houari Boummediene, Algiers, in 2001, 2003, and 2007, respectively. He was a full-time Researcher with the CERIST Research Center, Algeria, from 2002 to 2012. He is currently an Associate Professor with the Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia. His research interests include network security, intrusion detection systems, malware analysis, mobile security, and mobile networks.

SYED AHMAD CHAN BUKHARI (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of New Brunswick, Canada. Then went on to complete his postdoctoral fellowship at the Yale School of Medicine, where he worked with the Stanford University-Center of Expanded Data Annotation and Retrieval (CEDAR) to develop FAIR (Findable, Accessible, Interoperable, and Reusable) data submission pipelines to improve scientific experimental reproducibility. He is currently an Assistant Professor and the Director of Healthcare Informatics with St. John's University, New York City, NY, USA. His current research interests include addressing several core problems in the area of healthcare informatics and data science. He particularly focuses on devising techniques to semantically confederate heterogeneous biomedical data and to further develop clinical predictive models for diseases prediction. These techniques further alleviate many data access-related challenges faced by healthcare providers. He is also a Distinguished ACM Speaker who serves as an editorial board member of multiple scientific journals. His research work has published in top-tier journals and picked by various scientific blogs and international media.