Deep Learning for Diabetes: A Systematic Review

Taiyu Zhu, *Student Member, IEEE*, Kezhi Li, *Member, IEEE*, Pau Herrero, *Member, IEEE*, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*—Diabetes is a chronic metabolic disorder that affects an estimated 463 million people worldwide. Aiming to improve the treatment of people with diabetes, digital health has been widely adopted in recent years and generated a huge amount of data that could be used for further management of this chronic disease. Taking advantage of this, approaches that use artificial intelligence and specifically deep learning, an emerging type of machine learning, have been widely adopted with promising results. In this paper, we present a comprehensive review of the applications of deep learning within the field of diabetes. We conducted a systematic literature search and identified three main areas that use this approach: diagnosis of diabetes, glucose management, and diagnosis of diabetes-related complications. The search resulted in the selection of 40 original research articles, of which we have summarized the key information about the employed learning models, development process, main outcomes, and baseline methods for performance evaluation. Among the analyzed literature, it is to be noted that various deep learning techniques and frameworks have achieved state-of-the-art performance in many diabetes-related tasks by outperforming conventional machine learning approaches. Meanwhile, we identify some limitations in the current literature, such as a lack of data availability and model interpretability. The rapid developments in deep learning and the increase in available data offer the possibility to meet these challenges in the near future and allow the widespread deployment of this technology in clinical settings.**

***Index Terms*—Artificial intelligence, deep learning, deep neural networks, diabetes, diabetic complications, glucose management.**

I. Introduction

DIABETES is a group of lifelong metabolic disorders caused by defective insulin secretion or impaired insulin action. The International Diabetes Federation estimates that there are 463 million people (95% confidence interval: 369-601 million) living with diabetes in 2019, half of whom, however, remain undiagnosed, due to the complex pathogenesis of diabetes [1]. The global prevalence of diabetes is projected to significantly increase in the coming decade. Therefore, preventing and treating diabetes has been a heavy burden for national economies, healthcare systems, and personal medical expenditures, especially for low- and middle-income countries [2].

According to the etiopathology of diabetes, there are three main clinical categories: type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes mellitus (GDM) [3]. Other categories due to specific causes include latent autoimmune diabetes of adulthood and maturity-onset diabetes of the young. T1D occurs when the insulin-secreting β-cells of the pancreas are destroyed by the immune system [4]. People with T1D suffer from the absolute insufficiency of endocrine insulin produced by the pancreatic β-cell, hence they rely on exogenous delivery. T2D accounts for about 90% of people with diabetes, resulting from insulin resistance or insufficient insulin production. GDM appears during pregnancy and might require lifestyle interventions and exogenous insulin delivery to prevent complications in the infant. However, due to the increasing heterogeneity and lack of continuous monitoring, the early diagnosis and classification of diabetes are often difficult in practice [5].

The majority of people with diabetes needing exogenous insulin employ the so-called basal-bolus insulin therapy, which consists on measuring glucose levels with a glucose levels meter and delivering multiple daily injections (MDI) with an insulin pen or with an insulin pump (continuous subcutaneous insulin infusion (CSII)) [6].

For people living with diabetes, it is vital to maintain blood glucose (BG) levels in a normal range. Otherwise, hyperglycemia or hypoglycemia can cause short and long-term complications in microvascular and macrovascular, including neuropathy, nephropathy, retinopathy, stroke, cardiovascular disease, and peripheral vascular disease [7]. Nevertheless, BG control is challenging for people with diabetes, since there are plenty of daily factors that influence BG levels, such as meal ingestion, exercise, alcohol, illness, and stress. Thus self-management, e.g., timely BG measurement, hormone delivery, and adherence to recommended lifestyle are quite important, but all of them require multidisciplinary knowledge in clinical practice, especially for children and adolescents [8]. Besides, due to the high inter and intra-population variability in the glucose kinetics process and pharmacokinetics, it is difficult to find an optimal therapeutic strategy for all people [9].

Manuscript received June 12, 2020; revised October 22, 2020; accepted November 20, 2020. Date of publication November 24, 2020; date of current version July 20, 2021. This work was supported by EPSRC EP/P00993X/1 and President's Ph.D. Scholarship at Imperial College London. *(Corresponding author: Kezhi Li.)*

Taiyu Zhu, Pau Herrero, and Pantelis Georgiou are with the Centre for Bioinspired Technology, Imperial College London, SW7 2BU London, U.K. (e-mail: taiyu.zhu17@imperial.ac.uk; p.herrero-vinias@imperial.ac.uk; pantelis@imperial.ac.uk).

Kezhi Li is with the Institute of Health Informatics, University College London, WC1E 6BT London, U.K. (e-mail: ken.li@ucl.ac.uk).

Digital Object Identifier 10.1109/JBHI.2020.3040225

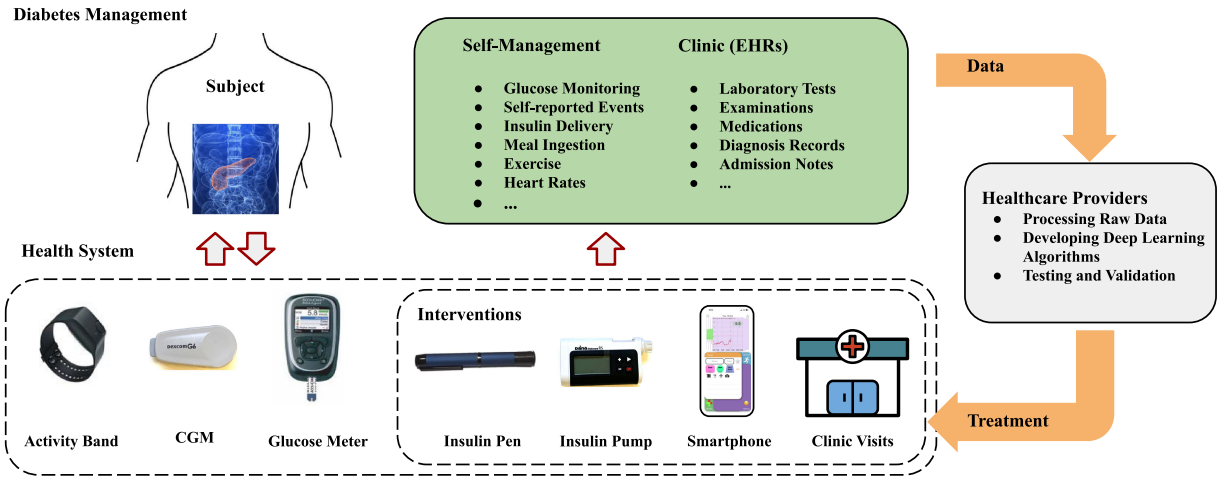

Fig. 1. Illustration of diabetes management, along with the large amount of data produced by self-management and clinical EHRs. The data is processed by healthcare providers to develop deep learning algorithms for the new treatments.

In recent decades, continuous glucose monitoring (CGM) systems [10]–[12] and closed-loop hormone delivery systems [13], [14], also known as the artificial pancreas (AP), have been widely researched, aiming at developing automatic glucose regulation and relieving the burden of glucose management. An AP system employs CGM, a closed-loop control algorithm, and an insulin pump to deliver insulin by CSII. It has been proven to effectively reduce glycaemic control and is recommended to some T1D cohorts [15]. Although the AP is currently the state of the art in insulin delivery, standard basal-bolus insulin therapy with a capillary blood glucose meter and MDI through an insulin pen remains a cost-effective treatment option, and in particular thanks to the recent improvements enabling wireless connectivity of these devices to a smartphone (i.e. smart pens and smart meters) have significantly enhanced this therapeutic option [16], [17]. Furthermore, significant progress has been made in developing smartphone applications and the integration of portable devices, such as CGMs, for diabetes care [18] which allow users to log their daily events and ultimately use them to provide decision support. There is an increasing interest in integrating physical activity monitoring through exercise bands to enhance BG management [19]. Consequently, the wide use of wearable devices and digital systems in diabetes management and the increasing electronic health records (EHRs) in clinics have produced a large amount of available data, as depicted in Fig. 1. This current scenario offers tremendous opportunities to apply advanced methods of artificial intelligence (AI) in diabetes care to further improve the treatment of diabetes.

The medical datasets from multiple sources are often heterogeneous, high-dimensional, and sparse, and thus they are likely to be underused in clinical scenarios [20]. Fortunately, machine learning, as an increasingly successful AI branch, is powerful at discovering nonlinear correlations of high-dimensional data. The definition of machine learning is that systems are able to learn knowledge and patterns automatically from experience or existing data without being explicitly instructed [21]. Moreover, empowered by boosting computational capabilities, a frontier machine learning method, deep learning, has achieved recent success and improved performance surpassing state-of-art in many health domains [22]. Compared to conventional machine learning technology, deep learning allows the input of raw data and learns representation automatically by exploiting deep neural networks (DNNs), which require little feature engineering work on data pre-processing [23].

Although there have been comprehensive literature reviews on AI for diabetes [24]–[26], covering some conventional machine learning methods and statistical models, they still lack a systematic study focusing on deep learning applications for diabetes. As an emerging approach, deep learning has recently shown competitive performance in several important fields of diabetes, such as diabetic eye diseases [27]. Therefore, in this work, we specifically investigate the latest advances in deep learning technologies for diabetes care.

II. Deep Learning Overview

Among the wide range of techniques and approaches in deep learning, we present the overview of several popular deep learning methods that are commonly applied to healthcare and, in particular, in the diabetes field. Deep learning originated from artificial neural networks (ANNs) inspired by the structure of biological neurons in the brain [28]. A standard ANN, as depicted in Fig. 2, comprises a number of nodes and three layers: input, hidden and output layer, to simulate the neuron behaviors by mathematical expressions. In general, an ANN gains the perceptions through an iterative training process called back-propagation but lacks generalization for supervised tasks [29]. By adding more hidden layers, deep learning extends the ANN structure to DNNs for better generalization, which extracts data features and learns representations with thousands or even millions of parameters [30]. The breakthroughs of computational hardware and software infrastructures largely accelerate the development of deep learning by increasing the size and depth of DNN models in the recent two decades [28]. Fig. 2 depicts

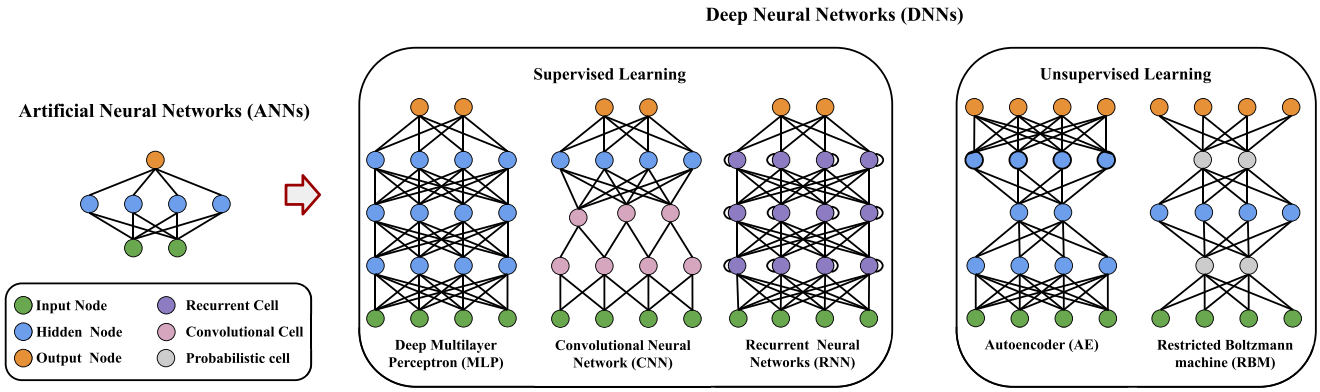

Fig. 2. Visualization of ANNs and DNNs. DNNs have an increasing number of hidden layers embedding the variants of neural nodes and cells. Higher-level feature maps are computed by the deep models. There are five popular DNN architectures for diabetes: DMLP, CNN, RNN, AE, and RBM.

five popular DNN architectures employed in diabetes research with the corresponding nodes, cells, and connections. Popular software frameworks to implement deep learning algorithms include Theano [31], Caffe [32], TensorFlow [33], CNTK [34], and PyTorch [35]. These frameworks support various programming languages and hardware acceleration, which help people efficiently build DNN models.

In general, most of the deep learning algorithms can be divided into supervised learning, unsupervised learning, and reinforcement learning. Classification and regression are common tasks in supervised learning, for which the labeled input data is used during iterative model optimization and backward propagation [29]. There are three supervised learning-based DNNs found in the literature of diabetes: deep multilayer perceptrons (DMLPs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). The DMLP, also known as a feed-forward neural network, uses the simple connections between neurons, i.e. fully connected (FC) layers, and forms the basis of many DNN models. The term "deep" is highlighted to indicate the modes have deep architectures with more than three layers since multilayer perceptrons refer to both ANNs and DNNs in some studies. A DMLP is associated with a set of weight vectors, bias scalars, and the nonlinear activation functions, including sigmoid, tanh, and rectified linear units (ReLU) [28].

Leveraging convolutional layers as preceptors, CNNs can process the signals of multi-dimensional arrays and achieve superior performance on imaging tasks [36]. A sub-sampling layer, or pooling layer, is employed in most CNN architectures to aggregate feature maps. One major advantage of convolutional operations is to reduce the neuron connections between layers, as depicted in Fig. 2, which notably enhances the efficiency of model training through back-propagation. Empowered by the parallelized operations of graphics processing units (GPUs) and tensor processing units (TPUs) [37], various CNN-based models have been applied to large-scale imaging recognition tasks, such as ImageNet database [38], and transformed to industry practices. In this regard, the popular CNN configurations found in literature are as follows: AlexNet [30], VGGNet [39], Inception (GoogLeNet) [40], and ResNet [41].

Different from other feed-forward neural networks, the input of an RNN contains the information at the previous timesteps. This feature makes RNNs powerful at processing sequential signals to capture temporal features. However, the difficulty of vanilla RNNs lies in the back-propagation training, where the gradient vanishing and exploding problems are likely to occur [42]. Fortunately, the advanced RNN cells, long short-term memory (LSTM) [43], and gated recurrent units (GRUs) [44], have overcome these problems by introducing gate functions and persevering long-term information. These RNN-based models have provided paradigms in numerous prediction and regression tasks, especially in natural language processing (NLP) and speech recognition. The latest trend in RNN is the attention mechanism [45], which allows models to focus on certain parts of input sequences and map the dependencies regardless of the distances.

Regarding unsupervised learning, the predefined labels or classes of inputs are not required for the model training. In this context, the algorithm aims at inferring the hidden structures and representations from the input datasets without supervision. Unsupervised learning is a powerful tool for data pre-processing, cluster analysis, density estimation, and dimension reduction. The autoencoder (AE) and the restricted Boltzmann machine (RBM) are the two basic architectures. The key feature of AEs is that its training target is the same as the input. Latent representations of the input are first transformed by an encoder, then fed to a decoder for reconstruction at the output. An RBM is another approach to map the representations by estimating probabilistic distribution over the input data, and thus it is also regarded as a generative model. Compared to standard Boltzmann machines, RBMs only allow the neuron connections that form a bipartite graph, to accelerate training processes. By stacking multiple RBMs, deep belief networks (DBNs) or deep Boltzmann machines can be constructed [46]. In most cases, DBNs are used as feature detectors to extract representations from data by unsupervised learning. However, supervised learning can be further performed to fine-tune the network weights and improve performance for certain learning tasks [47].

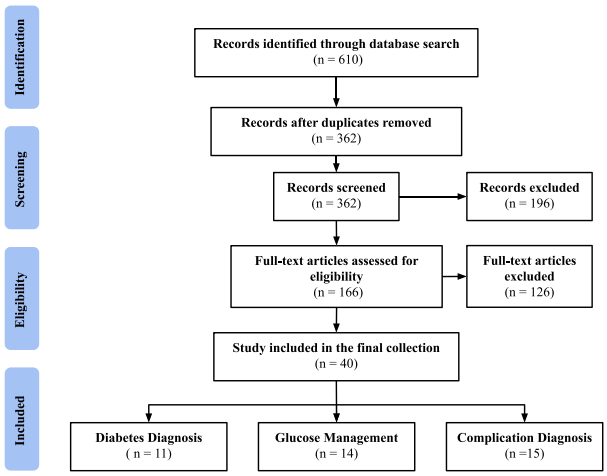

Fig. 3. PRISMA flow of selection process.

Deep reinforcement learning (DRL) surpass human professions in a variety of control problems with high-dimensional environments, where DNNs are employed as the approximators of policy, value-function, or system models. The network is trained by interacting with the environment consistently [48].

III. Methodology

Aiming at identifying and analyzing the benefits of deep learning within diabetes research, we conducted a systematic review by searching multiple public online databases, including PubMed, DBLP Computer Science Bibliography, and IEEEXplore. PubMed is a reputable database for biomedical and clinical research, while DBLP contains millions of publications in the field of computer science. IEEE Xplore is a digital library that covers studies in engineering and allied fields. All three databases provide free and open-access search engines or interfaces without requiring institutional subscriptions like other databases (e.g. Ovid, Scopus, and Web of Science). Therefore, to facilitate the reproducibility of the search results, we have chosen open-access search engines. We restricted the search to English-language documents that were published between January 1, 2016 and March 31, 2020 (first quarter, Q1). The search was performed between April 5 and May 1, 2020, and was based on titles, abstracts, and metadata. We followed preferred reporting items for systematic review and meta-analyses (PRISMA) approach [49]. Fig. 3 summarizes the selection process.

A. Search Strategies

In our paper search, the keywords "diabetes," "glucose" and "artificial pancreas" were combined with the deep learning terms using Boolean operators AND/OR. The specific query searched was: ((diabetes OR glucose OR artificial pancreas) AND (deep learning OR deep neural network OR convolution neural network OR convolutional neural network OR recurrent neural network OR LSTM OR autoencoder OR boltzmann machine OR deep belief network)). After obtaining the results of an initial collection of relevant articles, we first excluded duplicated articles from different sources, then performed a manual inspection to evaluate the remaining based on inclusion criteria.

B. Inclusion and Exclusion Criteria

The studies included in this review are original and available full-text, focusing on deep learning applications in diabetes. The final collection of articles was organized into three categories based on the clinical application: diagnosis of diabetes, glucose management, and diagnosis of complications. Particularly, the included studies were expected to:

1) present the details of datasets and data processing
2) explicitly describe methods, e.g., the structure of DNNs
3) evaluate model performance with standard metrics.

It should be noted that the application of diabetic retinopathy accounts for a large portion of the literature. Thus, in this area, we selected the works presenting DNN results of high originality, or large-scale clinical datasets. Abstracts, posters, technique reports, and reviews were excluded.

C. Information Extraction

From the selected collection of articles, we inspected the full-text and extracted key information to assess the deep learning applications. The following pre-defined categories were used to present the selected studies (Tables I, II, III).

1) Cases: We first summarized the specific application cases, i.e. scenarios, of the selected studies to identify the target of each work. For the studies with the available information of the types of diabetes, we have indicated them with † and ‡ for T1D and T2D, respectively.

2) Models: We present an overview of model architectures which includes a variety of DNN layers and the popular configurations, as mentioned in Section II. The details of hybrid structures and ensemble techniques are also included.

3) Data Sources: The source of input data is an essential factor for deep learning models. Many studies use more than one dataset, including public and private datasets, to validate the generalization of DNN models. Thus, this category summarizes the information regarding the employed datasets, e.g. sources, types, and formats. To facilitate future research to address the issues of data availability, we have highlighted the publicly available datasets with ⋆.

4) Development Process: This category summarizes the strategies for developing deep learning models, including preprocessing, training, validation, and testing. Although deep learning is good at extracting representations from raw data, these development steps need to be carefully designed, which impact on the functionality and reproducibility of the models.

5) Main Outcomes: The major outcomes with the corresponding metrics and criteria for performance evaluation, are included in this category. Some of the employed metrics in diabetes and complication diagnosis are sensitivity, specificity, and area under the curve (AUC); and root mean square error (RMSE) is common in glucose management. The results are consistent with the goals in the Cases category.

TABLE I
SUMMARY OF SELECTED ARTICLES FROM THE LITERATURE ON DIABETES DIAGNOSIS

Ref.	Cases	Models	Data Sources	Development Process	Main Outcomes	Baselines
[50]	Classification of diabetes‡	Denosing AE	Mount Sinai Data Warehouse⋆ (ICD-9)	Normalization; pre-process to obtain raw features; the data of training, validation and testing: 704,587, 5000, 76,214 patients	AUC: 0.907	Original descriptors, PCA (0.861)
[51]	Prediction of diabetes‡	Modified LSTM, attention pooling layer	An EHR dataset from a regional hospital (7191 patients, ICD-10)	The split for training, validation and testing: 2/3, 1/6 and 1/6 from 53,208 admissions	Precision of diagnosis, intervention, unplanned readmission: 66.2%, 78.7%, 79.0%	SVM, RF, plain RNN, LSTM (65.7%, 78.2%, 75.9%)
[52]	Detection of diabetes†‡	RBM and RNN	PID dataset from UCI repository⋆	Feature selection by RFs; min-max normalization; the ratio for training and testing data: 80%, 20%	Sensitivity and precision: 90.66%, 75%	N/A
[53]	Prediction of diabetes†‡	Modified 1-D CNN and FC layers	25 breath samples collected by MOS sensors with 1000-sec intervals	The data for training and testing: 15 samples, 10 samples; leave-one out cross-validation	AUC of T1D, T2D, healthy subjects: 0.9659, 0.9625, 0.9644	SVD, SVM, PCA
[54]	Detection of diabetes	5-layer CNN, LSTM, and SVM	ECG data sampled at 500 Hz with digital bandpass filtering and thresholding collected from 40 people	Heart rate variability (HRV) data from 71 ECG datasets (each contains 1000 samples); 5 fold cross-validation	Validation accuracy: 95.7%	Previous work using HRV
[55]	Detection of diabetes‡	DMLP with dropout	PID dataset from UCI repository⋆	The ratio of training and validation data: 90% and 10%	Accuracy: 88.41%	Previous work on the same dataset
[56]	Prediction of diabetes‡	DMLP	A population dataset (4814 participants,the majority are overweight)	Data cleaning (imputing missing values with the median); the ratio of training and testing data: 80% and 20% from 656 T2D subjects	AUC without and with HbA1c: 0.703, 0.840	SVM (0.679,0.825)
[57]	Prediction of the onset T2D‡	DMLP and a linear model	Practice Fusion dataset (9948 patients, ICD-9)⋆	Feature extraction by grouping 1312 features; the ratio of training and validation data: 70%, testing data: 30%; 10-fold cross-validation	Sensitivity: 31.17%, AUC: 84.13%	RF (29.12%, 16.07%)
[58]	Detection of diabetes‡	2 layer AE and a softmax layer	PID dataset from UCI repository⋆	Training the layer one by one with previous output; fine-tuning by supervised learning	Sensitivity: 87.92%, specificity: 83.41%, accuracy: 86.26%	Previous work on the same dataset
[59]	Prediction of diabetes‡	DBN	PID dataset from UCI repository⋆	Min-max normalization; feature selection by PCA; pre-training for RBMs; supervised fine-tuning	Sensitivity: 100%, F1 score: 0.808	DT, LR, RF, SVM, NB (75.9%, 0.760)
[60]	Detection of undiagnosed diabetes‡	2 hidden layer DMLP with dropout	An EHR dataset from a national survey (31,098 subjects, 4 years)	Combining 2013-2016 datasets;selecting features by LR; the data of training and testing: 11456 and 4444 subjects	AUC: 80.11%	LR, KNN, SVM, AdaBoost, Gaussian NB, RF (79.05%)

6) Baselines: In most selected studies, the authors implemented various baseline methods to compare with the performance of DNN algorithms. Many conventional statistic and machine learning methods are collected in this category, including logistic regression (LR), autoregression (AR), autoregressive integrated moving average (ARIMA), supporting vector machines (SVMs), random forests (RFs), naive Bayes (NB), k-nearest neighbors (KNN), latent variable model (LVX), principal component analysis (PCA), and decision trees (DTs). The best performance achieved by the baselines is also presented for the purpose of comparison using the metrics that are consistent with the Main Outcomes category.

7) Limitations: As a review for an emerging methodology such as deep learning, this category collects the limitations that were identified for the selected studies, which could inspire future work and improve the learning performance in each application area.

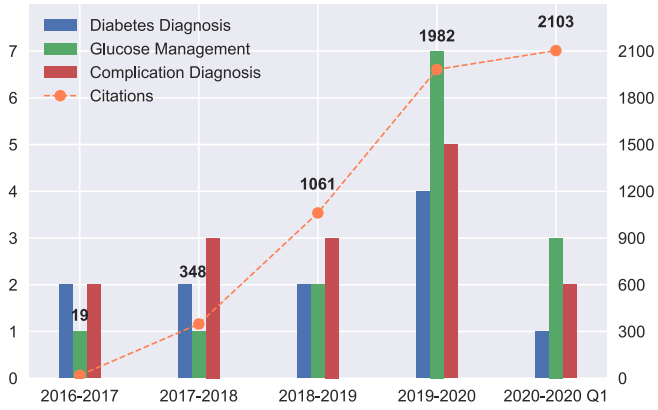

Fig. 4. Number of articles included in the collection grouped by the year of publication and application field. The orange dashed line indicates the number of citations in each year.

IV. RESULTS

The initial search yielded a total of 610 papers (PubMed (307), DBLP (31), and IEEE Xplore (272)), as shown in Fig. 3. After removing the duplicates, 362 papers remained. Then the papers were screened by the inclusion and exclusion criteria. We manually assessed the eligibility of the remaining papers by full-text inspection and included 40 papers in the final collection. Based on the application scenarios, we divided the final collection into three categories: diagnosis of diabetes (n = 11), glucose management (n = 14), and diagnosis of complications (n = 15). As shown in Fig. 4, most of the selected papers were published in recent two years, which indicates that the deep

TABLE II
SUMMARY OF SELECTED ARTICLES FROM THE LITERATURE ON GLUCOSE MANAGEMENT

Ref.	Cases	Models	Data Sources	Development Process	Main Outcomes	Baselines
[68]	Detection of hypoglycemia†	DBN by stacking RBMs	15 T1D children monitored for 10-hour overnight	Calculating QT correlation and heart rate; training and testing data: 10 and 5 subjects	Sensitivity: 79.70%, specificity: 50.00%	DBN based on ANNs (76.28%, 50.40%), multiple regression
[69]	Prediction of BG levels†	LSTM and a linear layer	A clinical database with the T1D data over 1600 days	Pre-training; linear interpolation; training and testing data: 5 patients in 400 days, another 5 patients with 200 samples	RMSE for 30, 60-min PH: 21.4, 38.0 mg/dL	ARIMA, EPM with SVR (21.6, 39.2 mg/dL)
[70]	Prediction of BG levels†	LSTM, Bidirectional LSTM and 3 FC layers	(1) GoCarb dataset (20 adults), (2) UVa/Padova T1D simulator (11 adults)	Pre-training on (2); linear interpolation on (1); cross-validation (67%, 33%); testing data: 26 sub-dataset from (1), 1791±141 CGM samples	RMSE for 30, 45, 60-min PH: 11.63, 21.75, 36.92 mg/dL	ARIMA, SVR (11.69, 22.14, 37.42 mg/dL)
[71]	Prediction of BG levels†	Deep sequential polynomial multi-output model (RNN)	40 T1D subjects over 1900 days	The ratio of training,validation and testing data: 85%, 7.5% and 7.5% from 555,000 CGM samples	Absolute percentage error for 30-min PH: 4.87	Linear extrapolation, RF, RNN(LSTM, 5.3)
[72]	Glycemic control†	CNN (Inception-v3)	Food-101 dataset (101 classes, 101,000 images)⋆	Image augmentation; pre-training on ImageNet dataset; training and testing data: 75,750, 25,250 images	Time in rage (TIR) of 70-180 mg/dL : 91.76%, top-1 accuracy of the image classification: 81.65%	Standard controller in UVa/Padova T1D simulator (TIR: 78.8%)
[73]	Prediction of BG levels‡	LSTM with dynamic time warping	A dataset from a randomized trial (26 adults, smartphone group (n = 11))	Pre-processing; transfer learning; min-max normalization across patients; training, validation and testing data: CGM samples of 120, 30 and 30days	Clark Error Grid zones of next-day PH (A: 84.12, B: 15.16, C: 0, D: 0.72, E: 0)%	ANN, KNN, ridge regression (A: 83.03%), kernel ridge regression, moving average
[74]	Prediction of BG levels†	CNN, LSTM and 2 FC layers	(1)UVa/Padova T1D simulator (10 adults), (2) 10 clinical subjects	Using Gaussian filter to remove outliers; training and testing data: 50% and 50%	RMSE for 30, 60-min PH: 9.38, 18.87 (1); 21.07, 33.27 (2) mg/dL	SVR (22.00, 34.35 (2)), LVX, neural network, AR
[75]	Prediction of BG levels†	LSTM and a FC layer	RT_CGM dataset (the population of 451 patients)	Removing sequences with low quality; Tikhonov regularization; training and testing data: 304,450 and 94128 samples	RMSE for 30, 45, 60-min PH: 19.47, 26.47, 32.38 mg/dL	AR, ANN, standard RNN, non-linear AR (24.66, 32.33, 38.58 mg/dL)
[76]	Prediction of BG levels†	Memory-Augmented LSTM with neural attention weights	(1) OhioT1DM dataset⋆, (2)AIDA Simulator, (3)UVa/Padova T1D simulator	Pre-training; linear interpolation and extrapolation; testing, validation and training data: last 10 days, previous 10 days, rest days (1); 400, 100, and 100 days (2); 70, 10 and 10 days (3)	RMSE for 30, 60-min PH: 18.74, 30.63 (1); 1.23, 2.27 (2); 2.93, 4.92 (3) with input of CGM, insulin and meal events	ARIMA (20.17, 33.47 (1); 5.59, 16.48 (2); 12.00, 18.66 (3))
[77]	Prediction of BG levels†	Dilated CNN (residual and parameterized skip connections)	(1) UVa/Padova T1D simulator, (2) ABC4D dataset, (3) OhioT1DM dataset⋆	Ruling out outliers; interpolation/extrapolation; label transformation; training, validation and testing set: 45%, 5%, 50% (1, 2); 40 and days (3)	RMSE for 30, 60-min PH: 8.88, 19.90 (1); 19.19, 31.78 (2); 19.28, 31.83 (3)	SVR (21.75, 34.31 (3)), LVX (12.25, 22.41 (1)), neural network (20.42, 33.13 (2)), AR
[78]	Glycemic control†	Deep Q-network with GRU or 1-D CNN	UVA/Padova simulator (30 virtual subjects)	Using CGM and insulin data in past 24 hours as the states, action setting: {0, basal rate, 5*basal rate}; testing in 10 days	Average risk index for the virtual subject: 9.26	Proportional-integral-derivative control (11.80)
[79]	Prediction of BG levels†	2 branches of LSTM cells (past and future information)	(1) UVa/Padova T1D simulator (100 adults), (2) Padova clinical dataset (1 patient)	Min-max normalization; output filtering in (2); training data: four-day protocol (1), testing data: 3-day scenario and *in vivo* data over a month (2)	Average RMSE for PH of 60 minutes: 11.72 (1), 21.09 (2)	Linearized average model (46.82 (2)), daily model predictor
[80]	Prediction of HbA1c†	1-D CNN, Inception module, FC layers	A clinical dataset (759 T1D subjects, 1543 observations)	Behavioral feature extraction; manual feature extraction; 10-fold cross-validation; batch normalization; loss regularization	Mean absolute error: 4.80, the coefficient of determination: 0.71	Nathan's formula, CNN (5.98, 0.62), manual features extraction network
[81]	Prediction of BG levels†	LSTM and 2 FC layers	OhioT1DM dataset⋆	Scaling glucose values by 0.01; the ratio of training, validation and testing data: 60%, 20% and 20%	RMSE for 30, 60-min PH: 18.867, 31.403	Previous work with machine learning

TABLE III
SUMMARY OF SELECTED ARTICLES FROM THE LITERATURE ON DIAGNOSIS OF COMPLICATIONS

Ref.	Cases	Models	Data Sources	Development Process	Main Outcomes	Baselines
[91]	Referable DR detection	CNN (Inspired by AlexNet, VGGNet)	(1) EyeCheck project, (2) Messidor-2 dataset⋆	Training data: 10,000 to 1,250,000 unique samples (1), testing data: 1748 images (2)	Sensitivity: 96.8%, specificity: 87.0%, AUC: 0.980	Classical detector by LR (AUC: 0.955))
[92]	Referable DR detection	CNN (Inception-v3), an ensemble of 10 networks	(1) EyePACS-1 dataset, (2) Messidor-2 dataset⋆	Batch normalization; pre-initialization by the ImageNet data; training and validation data : 80% and 20% of 128,125 retinal images, testing data: 4497 images (1) and 1748 images (2)	Sensitivity: 97.5% (1), 96.1% (2), specificity: 93.9% (1), 93.4% (2), AUC: 0.990 (1), 0.991 (2)	N/A
[93]	DR detection‡	A stack of non-negativity-constrained AEs	Images from 52 clinical scans with 12 retinal layers	Feature extraction by cumulative distribution function; training data: 40 subjects, testing data: 12 subjects	Sensitivity: 92%, specificity: 83%, accuracy: 100%	k-star (89%, 89%, 89%), KNN, RF, DT
[94]	DR detection	Customized CNN: (5 residual blocks), DT classifier	(1) EyePACS dataset, (2) Messidor-2 dataset⋆, (3) E-Ophtha dataset⋆	Feature extraction; training and validation data: 75,137 images (5-fold cross-validation, 1), testing data: 1368 images (2) and 405 images (3)	Sensitivity: 94% (1), 93% (2), 90% (3), specificity: 98% (1), 87% (2), 94% (3), AUC: 0.97 (1), 0.94 (2), 0.95 (3)	Previous work with machine learning on (2)
[95]	Referable DR detection	CNN (Adapted VGGNet)	Singapore national DR screening program	Each image was analyzed by two graders and one specialist, training data: 76,370 images (2010-2013 year), testing data: 71,896 images (2014-2015 year)	Sensitivity: 90.5%, specificity: 91.6%, AUC: 0.936	N/A
[96]	Referable DR detection	CNN (Inception-v3)	LabelMe dataset	21 grader validated the accuracy of the labels; training data: 58,790 images, cross-validation data: 8000 images	Sensitivity: 92.3%, specificity: 93.7%, 96% of participants satisfied with the model	Manual screening models
[97]	Moderate or worse DR detection	CNN (Inception-v4), an ensemble of 10 networks	EyePACS clinics, (2) Messidor-2 dataset⋆, (3) EyePACS-2 dataset	Gaussian process bandit algorithm (hyper-parameter tuning); training and validation data: 1,665,151 and 3737 images (1), (2), testing data: 1958 images (3)	Sensitivity: 97.1%, specificity: 92.3%, AUC: 0.986	Three retinal Specialists (sensitivity: 83.8%)
[98]	DR detection	CNN (VGGNet-s)	Kaggle dataset⋆	Normalization schemes and data augmentation; non-local means denoising; 5-fold cross-validation, training and validation data: 35,126 images	Sensitivity: 86.47%, specificity: 97.43%, AUC: 0.9786, accuracy: 95.68%	AlexNet, ResNet, VGGNet-16, VGGNet-19 (specificity: 96.49%), GoogleNet,
[99]	Referable DR detection	CNN (Inception-v4), an ensemble of 10 networks	A large-scale population (13 health regions, 7517 patients)	A cascade of thresholds (hyper-parameter tuning); testing data: 25,326 images	Sensitivity: 96.8%, specificity: 95.6%, AUC: 0.987	13 human regional graders (sensitivity: 74%)
[100]	Referable DR detection	CNN(Adapted VGGNet)	A multi-ethnic, multi-site dataset (5 races, 18,912 patients)	Training and validation data: 76,370 and 8000 images, testing data: 93,293 images	The estimation of DR prevalence: 16.1%, the AUC for referable DR: 0.863, the time taken to diagnose: 10.4h, risk factor: 0.743	10 retinal specialists and 7 professional graders (prevalence:15.9%, time: 1554.8 h)
[101]	Estimation of DR severity scale	CNN pillars (Inception-v3) and RF	(1) Kaggle dataset⋆, (2) 2 large clinical trials (530 patients)	SHAP for feature selection; transfer learning (1); 5-fold cross-validation: 4781 images (2)	AUC at month 6, 12, 24: 0.68, 0.79, 0.77	Well-trained reading center experts
[102]	Prediction of mortality in ICU	1-D CNN and 2 FC layers	MIMIC-III dataset⋆	Feature analysis by importance; addressing imbalance classes; training and testing data: 70% and 30% from 9000 subjects	AUC: 0.885	ANN (0.792), RF, DT
[103]	Prediction of myocardial infarction‡	DMLP	American commercial health plan	Descriptive statistics analysis, confounding factor analysis; extracting 199,116 patients	AUC: 0.767, with hazard ratio: 0.81 and 0.63	LR (AUC: 0.760)
[104]	Classification of diabetic foot	Customized 9-layer CNN	Plantar thermogram database with 167 subjects⋆	Data augmentation; patch extraction; the ratio of training and validation data: 70% and 30%, 10-fold cross-validation	Sensitivity: 0.9167, AUC: 0.8533	SVM, ANN, AlexNet, GoogLeNet
[105]	Detection of diabetic neuropathy†‡	U-Net CNNs (5 ensembles)	(1) BioImLab dataset⋆, (2) Beijing dataset, (3) ENA dataset	Patch extraction; training set: 1698 images (2), testing set: 2137 images (1), (3)	Fibre length 0.933, length/segment: 0.656, branch points: 0.891, nail points: 0.623	ACCMetrics model (0.825, 0.325, 0.570, 0.257)

learning research for diabetes is a fairly new topic and its interest has been accelerating. In addition, we also calculated and plotted the number of citations of the selected papers in October 2020, according to Google Scholar. The details of the selected works are presented in chronological order in Tables I, II, III.

A. Diagnosis of Diabetes

The early diagnosis of diabetes can effectively improve the medical care and treatment for people living with diabetes. The standard diagnosis to confirm diabetes in clinics requires repeated glucose-based tests on hemoglobin A1c (HbA1c) and corresponding diagnostic criteria for different diabetes types [3]. However, due to the huge population and shortage of physicians in rural areas, the number of undiagnosed cases is significant and projected to increase in the future [61]. There is a high risk of developing diabetes without the onset of symptoms, especially for people with T2D, which could lead to long-term dysfunction of various organs and chronic complications [62].

Therefore, the need to detect onset diabetes or predict the diabetes risk arises, e.g. population screening and non-invasive systems. Table I presents the current efforts at developing deep learning decision-support algorithms for the diagnosis of diabetes. In particular, various supervised and unsupervised learning approaches strategies have been applied, where DMLP models are the most widely employed. The feed-forward structures and simple connections make DMLP a good option for a binary classifier on EHRs, while AEs and RBMs are used to extract underlying patterns of the data without supervision. It is noted that many studies have used the publicly available dataset called Pima Indian Diabetes (PID), from the University of California Irvine (UCI) repository [63]. It contains 768 instances with eight attributes and a binary label (diabetic or non-diabetic), which can be visualized by [64]. The Pima Indians have a higher prevalence of T2D than any population [65], making the PID dataset popular in machine learning research. An advantage of using this dataset with the same metrics is to easily compare the results with the previous work employing various machine learning methods. In Table I, other public EHR datasets specific to the area of diagnosing diabetes include Mount Sinai Data Warehouse [66] and the Practice Fusion dataset [67]. Although these datasets are collected from different sources, they use common coding systems for diagnosis, namely the International Classification of Diseases ICD-9 and ICD-10. With these systems, researchers can easily locate the EHRs of patients with different types of diabetes and complications.

Nevertheless, the major limitation of applying deep learning on the PID dataset is the small number of patients and attributes. To prove the generalization of DNNs, the trained models need to be validated on a large population dataset. To this end, Miott *et al.* proposed a framework, namely Deep Patient, using a stack of denoising AEs to learn the representations from a large-scale dataset. The achieved AUC for the diagnosis of diabetes classification was 0.907 [50]. A recent study by Ryu *et al.* also employed a large dataset with 11 456 subjects [60]. They used a DMLP model to screen undiagnosed diabetes and achieved an AUC of 80.11% to detect undiagnosed diabetes. It is worth highlighting the data pre-processing step employed in these studies to extract related descriptors from the attributes of the patients, such as feature analysis and data normalization.

Moreover, the non-invasive detection of diabetes is emerging in several studies. Lekha *et al.* proposed a one-dimensional (1-D) CNN architecture to analyze the biomarkers in real-time breath signals for diabetes detection and classification [53]. The breath samples were collected by MOS sensors to analyze volatile organic compounds. The sensor array measured the content in a small gas chamber with an interval of 1000 seconds. Then the CNN classifier further processed these signals, which can reduce the need for feature selection and optimized the overall performance, compared to PCA, SVM, and singular value decomposition algorithm (SVD). In [54], the heart rate variability from electrocardiograms (ECG) was used as a marker to detect diabetes. The data was collected from a group of 40 people over a 10-minute duration. The ECG signals were sampled at 500 Hz with digital bandpass filtering and thresholding operations to reduce noise during the real-time detection. After deriving the heart rate time with the Pan-Tompkins algorithm, the study developed a hybrid deep learning model with CNN, LSTM, and SVM, and achieved a validation accuracy of 95.7%.

B. Glucose Management

The goal of glucose management in diabetes is to keep BG levels in the euglycemia region and avoid undesired glycemic events (i.e. hypoglycemia and hyperglycemia). Taking advantage of the digitization of diabetes self-management (Fig. 1), the development of deep learning has been significantly accelerated (Fig. 4). There are several sub-domains in glucose management which can be differentiated: BG level prediction, BG anomalies detection, insulin delivery control, and daily-life decision support.

Among these, BG level prediction has attracted increasing attention in recent years. An accurate BG prediction enables early interventions to prevent BG anomalies (i.e. hypoglycemia and hyperglycemia) and assists sensor-augmented insulin pumps (e.g. predictive low-glucose insulin suspend) and AP systems (e.g. model predictive control) to deliver optimal insulin and/or glucagon doses. The use of smartphone applications allows people to report the exogenous events that influence BG levels. By temporally aligning CGM measurements with these self-reported events, such as meal composition and insulin dosage, a multivariate time series can be formed and processed by deep learning models. Normally, the prediction horizon (PH) for short and long-term forecasting is 30 minutes and greater than 60 minutes respectively. In this scenario, the RNN-based architecture is a powerful tool, referring to its success in temporal sequence processing and regression. Augmented by LSTM cells, the RNN is the most widely used method for glucose prediction in Table II. Mirshekarian *et al.* proposed an LSTM model for 30 and 60-minute prediction, which outperforms the engineered physiological model (EPM) with SVR [69]. EPM is a continuous dynamic model used to calculate the system states, which comprises the compartments of meal absorption dynamics, insulin absorption dynamics, and glucose-insulin

dynamics. They further introduced a neural attention layer to emulate the case-based prediction by a memory module [76]. Besides, Li *et al.* transferred the prediction into a classification problem and used 1-D dilated CNNs in their GluNet framework to classify the predictive changes of future BG values [77]. The study tested the model on the two clinical datasets. The use of dilated DNNs to improve the BG prediction is also highlighted in [82]–[84]. Similarly, Zaitcev *et al.* employed 1-D CNNs with the Inception module to estimate HbA1c from imperfect time series of CGM [80]. An important consideration for BG prediction is whether the algorithm can be applied in real-time. In this regard, the deep learning models based on CNN and LSTM layers have been validated in smartphone applications to perform real-time BG prediction with short inference time and small memory consumption [74], [77].

Another application of CNNs is to estimate macronutrient content as a daily-life support [72]. With a publicly available dataset of food images (Food-101 dataset) [85], the trained CNN model can predict the food category based on the food images from smartphones, then assist decision support systems and AP systems to compute the required amount of meal bolus insulin. The proposed algorithm was validated in the UVA/Padova T1D simulator with the disturbances of carbohydrate content and incorrect estimation of meal sizes. The UVA/Padova T1D simulator, developed by the University of Virginia (US) and the University of Padova (Italy), is a glucose-insulin dynamics simulator that has been accepted by the Food and Drug Administration for pre-clinical studies [86]. In recent years, many research groups use computer simulation, i.e. *in silico* setup, to test algorithms in various virtual scenarios, considering the high costs and safety concerns associated to actual clinical trials in humans and animals. Fox *et al.* used the UVA/Padova T1D simulator to test DRL algorithms to control the delivery of basal insulin, using GRU and 1-D CNN architectures. Recent studies also explored the latest DRL algorithms for other types of hormones in glycemic control, such as glucagon and bolus insulin [87], [88]. Moreover, the simulator is frequently employed in glucose prediction to generate synthetic population datasets that are used for initial validation. Although most of the studies conducted experiments on their proprietary clinical datasets, there is a dataset available to researchers, the OhioT1DM dataset, which was released for the first edition of BG level prediction challenge in 2018 and later updated for the 2020 edition [89]. This dataset contains multi-modal data (CGM, meals, insulin, exercise) corresponding to 12 subjects with T1D over eight weeks.

Furthermore, an unsupervised learning algorithm based on DBNs, and taking ECG signals as input, was employed to detect hypoglycemia in children with T1D [68]. Similarly, in a recent study, ECGs were used to detect nocturnal hypoglycemia in healthy individuals with a CNN-LSTM model [90].

C. Diagnosis of Complications

In this category, most research has focused on the analysis of medical imaging to detect and diagnose multiple complications associated to people living with diabetes, as shown in Table III. Diabetes-related complications are diverse and regular examinations and clinical visits are time-consuming, expensive, and subjective [106]. For a long-term chronic disorder such as diabetes, its treatment is a heavy burden on the healthcare systems. Therefore, automated systems that are able to screen, detect, predict, and diagnose diabetes-related complications play an important role in population-based surveillance and monitoring.

Diabetic retinopathy (DR) is the leading cause of vision impairments and blindness in the world [107]. DR is often difficult to be detected until vision-threatening events occur. Fortunately, the state-of-art technologies of deep learning have shown great potential to meet this challenge and provide solutions to various DR problems reaching, in some cases, superhuman performance [27]. Following the success in the computer version (CV), a large number of CNN-based models have been adopted to extract the representation from retinal fundus photographs. In the 2015 Kaggle competition on DR screening, all the top results were achieved by CNNs, using a publicly available dataset [108]. Other public datasets with images of DR examinations include the Messidor-2 dataset [109] and the E-Ophtha dataset [110]. In Table III, nearly all the selected studies used CNNs to detect DR (10/11, 91%). The exception is the study by ElTanboly *et al.* which designed a multistage deep fusion classification network with a stack of non-negativity-constrained AEs to detect DR in optical coherence tomography (OCT) images for the patients who have almost normal retina appearances [93]. The AE model achieved high classification accuracy on an experiment with 52 subjects. As for CNN-based studies, most of the approaches are adapted or inspired from two popular architectures in the CV; VGGNet (4/10, 40%) and Inception (5/10, 50%). In [98], multiple popular CNN configurations were explored on the Kaggle dataset, where VGGNet-s obtained the highest accuracy in the experiments. VGGNet was developed by the University of Oxford (U.K.), aiming at improving the recognition performance on the ImageNet database with small kernel size and deep networks [39], while Inception employs sparse connections between activation functions in an Inception module to enhance the efficiency of computation on GPUs [40]. Both of the architectures achieved satisfactory performance on DR detection. Abràmoff *et al.* proposed a VGGNet-based model to detect multiple classes of DR on Messidor-2 dataset [91], achieving a high sensitivity of 96.8% on referable DR. Then the VGG adapted architecture was validated in two large scale datasets with multi-ethnic populations [95], [100]. Their studies indicated deep learning methods can detect the referable DR with high accuracy but with much less time than human assessors. Gulshan *et al.* used an Inception-based architecture to detect referable DR and achieved the sensitivity of 96.8% and the specificity of 87.0%. The clinical settings to implement such systems were further investigated, including the feasibility and acceptability of outpatient settings [96] and grader variability [97]. Ruamviboonsuk *et al.* conducted a nationwide experiment to validate an Inception-based model [99]. Compared with human specialists, the deep learning model obtained significantly higher sensitivity and slightly lower specificity. Their achievement is regarded as one of the human level performance milestones in the AI Index 2019 annual report [111]. Meanwhile, Arcadu *et al.*

proposed an Inception model to predict DR progression by leveraging individual color fundus photographs [101].

The deep learning applications to other complications are also noted in the literature. Wittler *et al.* designed a CNN-based model to predict mortality based on the data from the intensive care unit (ICU) patients, achieving an AUC score of 0.885 [102]. The ICU dataset in this study, referred to as MIMIC-III, is freely accessible [112]. Williams *et al.* [105] proposed a U-Net CNN to quantify the nerve fiber properties in the diagnosis of diabetic neuropathy, involving a public dataset [113]. Their results show an excellent localization performance for the quantification and the potential to be adopted in clinical settings. In [104], a customized CNN was designed to detect plantar ulcers on the thermography of diabetic foot with a publicly accessible dataset [114]. Moreover, Yamada *et al.* investigated the incidence of cardiovascular disease among three anti-diabetic drugs, using a DMLP model that achieved better results than conventional LR analysis [103]. In summary, it is noted that most of the studies focused on microvascular complications, including DR [91]–[101], diabetic foot [104], and diabetic neuropathy [105], while there is only one study focusing on macrovascular complications (cardiovascular diseases) [103].

D. Summary of Deep Learning Techniques

A significant number of deep learning methods have been adopted by the diabetes research community, covering various architectures in supervised learning and unsupervised learning. Among these, CNN-based architectures are the most widely used, particularly in clinical imaging problems. CNNs are good at extracting features from raw data, requiring little hand-engineered work and domain expertise on image processing [23]. Hence, the main application of CNNs has been in the analysis of clinical scans and medical images used for the diagnosis of diabetes-related complications. Another use of CNN-based architectures is to support the daily life of people with diabetes by estimating macronutrients from food images. Today the new techniques are actively developed in the CV area to improve the model performance while reducing the complexity. Hence the evolution of CNN configurations (VGGNet, Inception) opens the door to the development of more powerful algorithms. In addition, some studies have applied 1-D CNNs to process sequential signals, using convolutional filters to extract data features with a large receptive field. Several of these work further used LSTM layers to process the outcomes from CNN architectures to compute the temporal dependencies by hybrid models: convolutional RNN (CRNN) [74] and CNN-LSTM [54], [90], relying on the powerful capability of RNNs in sequence processing. In fact, RNN-based architectures dominate among the glucose management applications, especially for BG prediction. The recursive computations and advanced cell structures are suitable for mapping the glucose series measured by CGM in real-time. Some studies explored the latest advances of the techniques in NLP, such as bidirectional LSTM [70] and neural attention mechanisms [51], [76].

DMLP and unsupervised learning algorithms are commonly used in diabetes diagnosis. However, careful feature selection and normalization in pre-processing are required in many cases of these tasks, due to the heterogeneous forms of the records in EHR datasets. To this end, conventional machine learning algorithms are employed to discover the most relevant features. For instance, in [59], PCA was employed to calculate principal components scores for each data feature by deriving the eigenvectors coefficients and weights. In [60], LR analysis was used to compute correlations between non-invasive variables and the attributes of individuals to select significant features for diabetes detection.

Aiming at specific tasks, the DNN layers can be accessed and customized to incorporate with other models. Apart from the CRNN and the CNN-LSTM, other hybrid learning models in the literature involve the linear model [57], SVM [54], DTs [94], and RFs [101] to integrate data features at the input or perform a second-level analysis at the output. Moreover, the ensemble models for deep learning are highlighted in [92], [97], [99], [101], [105]. In these works, the ensemble contains multiple CNNs trained by the same dataset and obtains the final results by linear averaging in the testing phase. Due to the random initialization and batch feeding, each of the CNNs learns a distinct representation and improves the overall accuracy and generalization.

Training a very deep model from scratch is time-consuming since millions of parameters in the DNN units need to be tuned. In this regard, an approach called transfer learning, i.e. pre-training, provides a shortcut to solve this issue. Particularly, for clinical imaging tasks, the ImageNet database is a critical auxiliary component for CNN pre-training, which has been used in [72], [92]. Fine-tuning the weights based on the ImageNet paradigms can largely speed up convergence for target datasets, but it is not a necessary step if there are sufficient computational resources [115]. In the tasks of glucose management, the *in silico* datasets derived from the simulators and a portion of real clinical data are used for pre-training [69], [70], [76]. It is an effective method to mitigate the high demand for data during the DNN training. The unsupervised pre-training in DBNs is calculated layer by layer [47] to find good initial weights for the discriminative fine-tuning, which is used by [59]. Alternatively, data augmentation is a way to improve model performance with limited data, which can be found in [98], [104]. These studies used a series of image manipulation, such as shifting, rotating, and flipping, to transform the available images and expand the training sets.

V. Discussion

A. Limitations and Challenges

Although deep learning has improved the state of art in several areas of diabetes, the applications in healthcare systems need to be robust, reliable, and convincing to avoid safety issues and provide effective therapeutic aids. In this context, there remain several limitations and challenges for deep learning to be further introduced in actual clinical settings. Table IV summarizes the five common limitations identified from the selected articles: data volume, data variability, data quality, feature processing, and interpretability. In real-world scenarios,

TABLE IV
SUMMARY OF THE LIMITATIONS AND CHALLENGES IDENTIFIED BY THE SELECTED ARTICLES

Category	Description	References
Data volume	Training a deep model for complicated tasks requires a high volume of data. Collecting data from people with diabetes is often time-consuming and expensive, compared to other tasks in CV and NLP. Consequently, many studies face a shortage of data during their research cycles.	[51], [53], [54], [56], [73], [81], [91], [94], [98], [101], [104]
Data variability	The variability among people with diabetes is large due to the complex glucose dynamics. To obtain better generalization for deep learning models, the training data needs to cover a diverse range of individuals, such as people of different ages and comorbidities. However, many datasets are often collected from a specific cohort of people, which lacks diversity and could bring bias to the learning.	[50], [51], [53], [56], [60], [68], [72], [73], [75], [80], [92], [96], [97], [99]–[101], [103]
Data quality	Similar to many other problems in healthcare, most of the diabetes datasets are heterogeneous, sparse, and noisy with some missing values. It is not realistic to collect perfect data from either clinical practice or daily self-management, e.g. the unavoidable errors from CGM sensors.	[50], [53], [57], [73], [77], [78], [80], [81], [91], [102]
Feature processing	The major challenge in feature processing is to find the most effective features for models to learn the representations. Manually screening and analyzing each feature in a diabetes dataset could require a lot of engineering work, but using automated data-driven methods, such as PCA, would ignore some physiological knowledge and rely too much on the characteristics of the data. A more comprehensive analysis of additional factors and features is needed with the advances in data collecting and physiological models.	[57], [69], [70], [72], [74], [77], [92], [98], [100], [103]
Interpretability	The interpretability, i.e. explainability, stands for how the model obtains the corresponding output based on a set of inputs. It is an important goal for AI applications in healthcare to convince clinicians to adopt such systems. In many cases, deep learning models are regarded as "black boxes" with a lack of model transparency due to complex nonlinear layers. As a consequence, if the model performance degrades in certain circumstances, it might be difficult to explain why.	[74], [91], [92], [94], [95], [99], [101]

the data collected from people with diabetes are prone to be imperfect, due to human errors and sensor artifacts. The process to collect real data is sometimes expensive and time-consuming. Due to data privacy policies, sharing data-sets among research teams is sometimes difficult. These factors lead to many studies employing a reduced, sometimes insufficient, amount of data. Another challenge that arises due to the complexity of glucose dynamics is how to process the available data in order to characterize people with diabetes. Also, deep learning models lack transparency. From the perspective of clinicians, why the models produce the output for a certain input case is important, particularly for some critical decision-making applications. The complicated structures in DNN layers can effectively learn the patterns from non-linear signals but reduce the interpretability of the model. Therefore, it is crucial to consider the trade-off between performance and interpretability when investigating deep learning for diabetes. Finally, the efficiency of training deep learning models is expected to be enhanced through new algorithmic and hardware developments [52], [104].

B. Opportunities and Future Work

The list of challenges introduced in Section V-A not only applies to the field of diabetes but also is valid in other health domains. Deep learning is a hotspot in the era of AI, and it is worth noting that most of the selected papers are publications from the recent two years, as shown in Fig. 4, which indicates that this is an emerging technology. Hence, there is a large space to improve the current applications for diabetes.

First, the digital records and vital signs are increasingly collected by the multi-modal systems with wearables and smartphone applications. Most of these data are conveniently uploaded to centralized systems or cloud repositories. With the popularization of the Internet of things and 5 G networks, data volumes and variability of data sources are expected to significantly increase in many healthcare applications, and in particular, in diabetes care. As the data volume expands, many low-quality data samples can be filtered out and removed from training sets, and the advances in wearables (e.g. CGM) can effectively reduce the measurement errors. Deep learning is well adapted to cope with such an increase in data availability. Several publicly available datasets are outlined in Section IV, and more datasets will be shared in the communities after proper post-processing and de-anonymization. In order to deploy deep learning in an ambulatory setting, the frameworks mentioned in Section II can be easily ported to mobile devices by using tools such as TensorFlow Lite [74], [77].

To interpret deep learning technologies in healthcare, many recent attempts in the AI domain have been made to enhance model transparency and understand model functionality. In particular, a unified framework, the SHapley Additive exPlanations (SHAP), was proposed to explain the input features that contribute to the final output, which has been validated on many data-driven applications in healthcare domains [116]. This is also an effective method to select input features by ranking their importance. In Table III, an article also employed SHAP analysis to attribute the descriptors for the CNN outcomes [101]. Another effective technique to interpret the learned features of CNN layers is t-distributed stochastic neighbor embedding (t-SNE) [117], which was used to visualize the clusters of heartbeat data according to glucose levels in [90]. The use of t-SNE can also be generalized to other CNN applications, such as DR detection, to qualitatively analyze the extracted feature maps. Moreover, a recent study also verified the conformance of neural network models in terms of glucose-insulin dynamics [118]. Similar approaches can be used to analyze the performance of DNNs and further enhance interpretability.

Instead of solely using data-driven models, integrating the expert knowledge in the learning process can help to better understand the underlying mechanisms of a health condition such as diabetes. Specifically, there are two feasible methods. One is to incorporate the physiological parameters as the input feature of the models, and the other is to use expert knowledge as a guide during the training process. Expert knowledge is also

essential to craft safety constrains and calculate the confidence of the model outputs.

Many selected articles mentioned that their studies require to be further validated in real-world scenarios [79], [80], [95], [101], [105]. In this regard, a team from Google took a step forward. They conducted a human-centered study in 11 clinics, applying deep learning to diabetic eye diseases [119]. The results indicate that some socio-environmental factors need to be addressed before the widespread deployment of such automated systems.

VI. Conclusion

In this paper, we present a comprehensive review of the current trend in deep learning technologies for diabetes research. We performed a systematic search, selected a collection of articles, and summarized the key information focusing on three areas: diagnosis of diabetes, glucose management, and diagnosis of diabetes related complications. In these areas, various DNN architectures and learning techniques have been applied and obtained superior experimental performance that previous conventional machine learning approached. On the other hand, several challenges have been identified from the literature including data availability, feature processing, and model interpretability. In the future, there is great potential to meet these challenges by transferring the latest advances in deep learning technologies into massive multi-modal data of diabetes management. We expect that deep learning technologies will be widespread in clinical settings and largely improve the treatment of people living with diabetes.

References

[1] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas," *Diabetes Res. Clin. Pract.*, vol. 157, 2019, Art. no. 107843.

[2] *Global Report Diabetes*. World Health Organization, 2016.

[3] American Diabetes Association and others, "2. classification and diagnosis of Diabetes," *Diabetes Care*, vol. 40, no. Supplement 1, pp. S11–S24, 2017.

[4] Centers for Disease Control and Prevention, "National diabetes fact sheet: National estimates and general information on Diabetes and prediabetes in the U.S. 2011," U.S. Dept. Health and Human Services, 2011.

[5] T. Tuomi, N. Santoro, S. Caprio, M. Cai, J. Weng, and L. Groop, "The many faces of diabetes: A disease with increasing heterogeneity," *Lancet*, vol. 383, no. 9922, pp. 1084–1094, 2014.

[6] J. C. Pickup, "Management of diabetes mellitus: Is the pump mightier than the pen?," *Nat. Rev. Endocrinol.*, vol. 8, no. 7, 2012, Art. no. 425.

[7] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, "Epidemiology of diabetes and diabetes-related complications," *Phys. Ther.*, vol. 88, no. 11, pp. 1254–1264, 2008.

[8] R. D. Coffen and L. M. Dahlquist, "Magnitude of type 1 diabetes self-management in youth health care needs diabetes educators," *Diabetes Educator*, vol. 35, no. 2, pp. 302–308, 2009.

[9] J. Vora and T. Heise, "Variability of glucose-lowering effect as a limiting factor in optimizing basal insulin therapy: A review," *Diabetes, Obesity Metab.*, vol. 15, no. 8, pp. 701–712, 2013.

[10] D. C. Klonoff, "Continuous glucose monitoring: roadmap for 21st century diabetes therapy," *Diabetes Care*, vol. 28, no. 5, pp. 1231–1239, 2005.

[11] Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group, "Continuous glucose monitoring and intensive treatment of type 1 Diabetes," *New Engl. J. Med.*, vol. 359, no. 14, pp. 1464–1476, 2008.

[12] D. Rodbard, "Continuous glucose monitoring: A review of successes, challenges, and opportunities," *Diabetes Technol. Ther.*, vol. 18, no. S2, pp. S2–S3, 2016.

[13] P. Herrero *et al.*, "The bio-inspired artificial pancreas for type 1 diabetes control in the home: System architecture and preliminary results," *J. Diabetes Sci. Technol.*, vol. 13, no. 6, pp. 1017–1025, 2019.

[14] C. Fabris and B. Kovatchev, "The closed-loop artificial pancreas in 2020," *Artif. Organs*, vol. 44, no. 7, pp. 671–679, 2020.

[15] A. Slomski, "Artificial pancreas improves glycemic control," *JAMA*, vol. 320, no. 20, pp. 2068–2068, 2018.

[16] T. S. Bailey and J. Y. Stone, "A novel pen-based bluetooth-enabled insulin delivery system with insulin dose tracking and advice," *Expert Opin. Drug Del.*, vol. 14, no. 5, pp. 697–703, 2017.

[17] M. Grady, L. B. Katz, H. Cameron, and B. L. Levy, "Diabetes app-related text messages from health care professionals in conjunction with a new wireless glucose meter with a color range indicator improves glycemic control in patients with type 1 and type 2 Diabetes: Randomized controlled trial," *JMIR Diabetes*, vol. 2, no. 2, 2017, Art. no. e19.

[18] L. F. Garabedian, D. Ross-Degnan, and J. F. Wharam, "Mobile phone and smartphone technologies for diabetes care and self-management," *Curr. Diabetes Rep.*, vol. 15, no. 12, 2015, Art. no. 109.

[19] F. L. Schwartz, C. R. Marling, and R. C. Bunescu, "The promise and perils of wearable physiological sensors for diabetes management," *J. Diabetes Sci. Technol.*, vol. 12, no. 3, pp. 587–591, 2018.

[20] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nat. Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[22] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, 2016.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] M. Rigla, G. García-Sáez, B. Pons, and M. E. Hernando, "Artificial intelligence methodologies and their application to diabetes," *J. Diabetes Sci. Technol.*, vol. 12, no. 2, pp. 303–310, 2018.

[25] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: Literature review," *J. Med. Internet Res.*, vol. 20, no. 5, 2018, Art. no. e10775.

[26] I. Dankwa-Mullan, M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, and K. Rhee, "Transforming diabetes care through artificial intelligence: The future is here," *Population Health Manag.*, vol. 22, no. 3, pp. 229–242, 2019.

[27] A. Grzybowski *et al.*, "Artificial intelligence for diabetic retinopathy screening: A Review," *Eye*, pp. 1–10, 2019.

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learn.*. Cambridge, MA, USA: MIT press, 2016.

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[31] J. Bergstra *et al.*, "Theano: A CPU and GPU math expression compiler," in *Proc. Python Sci. Comput. Conf. (SciPy)*, vol. 4, no. 3, Austin, TX, 2010, pp. 1–7.

[32] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[33] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc.12th {USENIX} Symp. Operating Syst. Des. Implementation*, 2016, pp. 265–283.

[34] F. Seide and A. Agarwal, "CNTK: Microsoft's open-source deep-learning toolkit," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 2135–2135.

[35] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[37] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural Comput.*, vol. 22, no. 12, pp. 3207–3220, 2010.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations,* Y. Bengio and Y. LeCun, Eds. 2015, p. 1.

[40] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, 1994.

[43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[44] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[46] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, pp. 5947–5947, 2009.

[47] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[48] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[49] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Internal Med.*, vol. 151, no. 4, pp. 264–269, 2009.

[50] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, 2016.

[51] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Inform.*, vol. 69, pp. 218–229, 2017.

[52] S. Ramesh, H. Balaji, N. C. S. Iyengar, and R. D. Caytiles, "Optimal predictive analytics of Pima diabetics using deep learning," *Int. J. Database Theory Appl.*, vol. 10, no. 9, pp. 47–62, 2017.

[53] S. Lekha and M. Suchetha, "Real-time non-invasive detection and classification of diabetes using modified convolution neural network," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1630–1636, 2017.

[54] G. Swapna, R. Vinayakumar, and K. Soman, "Diabetes detection using deep learning algorithms," *ICT Exp.*, vol. 4, no. 4, pp. 243–246, 2018.

[55] A. Ashiquzzaman *et al.*, "Reduction of overfitting in diabetes prediction using deep learning neural network," in *IT Convergence Secur. 2017*. Berlin, Germany: Springer, 2018, pp. 35–43.

[56] S. Spänig, A. Emberger-Klein, J.-P. Sowa, A. Canbay, K. Menrad, and D. Heider, "The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes," *Artif. Intell. Med.*, vol. 100, 2019, Art. no. 101706.

[57] B. P. Nguyen *et al.*, "Predicting the onset of type 2 Diabetes using wide and deep learning with electronic health records," *Comput. Methods Prog. Biomed.*, vol. 182, 2019, Art. no. 105055.

[58] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 Diabetes data classification using stacked autoencoders in deep neural networks," *Clin. Epidemiol. Global Health*, vol. 7, no. 4, pp. 530–535, 2019.

[59] P. Prabhu and S. Selvabharathi, "Deep belief neural network model for prediction of Diabetes Mellitus," in *Proc. 3rd Int. Conf. Imag., Signal Process. Commun.*. 2019, pp. 138–142.

[60] K. S. Ryu, S. W. Lee, E. Batbaatar, J. W. Lee, K. S. Choi, and H. S. Cha, "A deep learning model for estimation of patients with undiagnosed diabetes," *Appl. Sci.*, vol. 10, no. 1, 2020, Art. no. 421.

[61] N. Cho *et al.*, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, 2018.

[62] American Diabetes Association and others, "Screening for diabetes," *Diabetes Care*, vol. 25, no. suppl 1, pp. s21–s24, 2002.

[63] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[64] R. Rossi and N. Ahmed, "The network data repository with interactive graph analytics and visualization," in *Proc. 29th AAAI Conf. Artif. Intell.*, Jan. 2015, pp. 4292–4293.

[65] R. S. Lindsay *et al.*, "Adiponectin and development of type 2 Diabetes in the Pima Indian population," *Lancet*, vol. 360, no. 9326, pp. 57–58, 2002.

[66] Icahn School of Medicine at Mount Sinai, "Mount Sinai Data Warehouse," 2015. [Online]. Available: https://labs.icahn.mssm.edu/msdw/

[67] Practice Fusion, Inc, "Practice Fusion." 2020. [Online]. Available: https://www.practicefusion.com/

[68] P. P. San, S. H. Ling, and H. T. Nguyen, "Deep learning framework for detection of hypoglycemic episodes in children with type 1 Diabetes," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 3503–3506.

[69] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz, "Using LSTMs to learn physiological models of blood glucose behavior," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 2887–2891.

[70] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network," in *Proc. 14th Symp. Neural Netw. Appl. (NEUREL)*, 2018, pp. 1–5.

[71] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, and J. Wiens, "Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1387–1395.

[72] A. Chakrabarty, F. J. Doyle, and E. Dassau, "Deep learning assisted macronutrient estimation for feedforward-feedback control in artificial pancreas systems," in *Proc. Annu. Amer. Control Conf.*, 2018, pp. 3564–3570.

[73] S. H. A. Faruqui *et al.*, "Development of a deep learning model for dynamic forecasting of blood glucose level for type 2 diabetes mellitus: Secondary analysis of a randomized controlled trial," *JMIR mHealth uHealth*, vol. 7, no. 11, 2019, Art. no. e14452.

[74] K. Li, J. Daniels, C. Liu, P. Herrero-Vinas, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE J. Biomed. Health Infrom.*, vol. 24, no. 2, pp. 603–613, Feb. 2020.

[75] A. Aliberti, I. Pupillo, S. Terna, E. Macii, S. Di Cataldo, E. Patti, and A. Acquaviva, "A multi-patient data-driven approach to blood glucose prediction," *IEEE Access*, vol. 7, pp. 69 311–69 325, 2019.

[76] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "LSTMs and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2019, pp. 706–712.

[77] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting," *IEEE J. Biomed. Health Infrom.*, vol. 24, no. 2, pp. 414–423, Jul. 2019.

[78] I. Fox and J. Wiens, "Reinforcement learning for blood glucose control: Challenges and opportunities," in *Proc. Reinforcement Learn. Real Life Workshop 36th Int. Conf. Mach. Learn.*, 2019, pp. 1–8.

[79] E. M. Aiello, G. Lisanti, L. Magni, M. Musci, and C. Toffanin, "Therapy-driven deep glucose forecasting," *Eng. Appl. Artif. Intell.*, vol. 87, 2020, Art. no. 103255.

[80] A. Zaitcev, M. R. Eissa, Z. Hui, T. Good, J. Elliott, and M. Benaissa, "A deep neural network application for improved prediction of HbA1c in type 1 Diabetes," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2932–2941, Oct. 2020.

[81] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, "Blood glucose prediction with variance estimation using recurrent neural networks," *J. Healthcare Informat. Res.*, vol. 4, no. 1, pp. 1–18, 2020.

[82] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction," in *3rd Int. Workshop Knowledge Discovery Healthcare Data, IJCAI-ECAI 2018*, 2018, pp. 64–78.

[83] J. Chen, K. Li, P. Herrero, T. Zhu, and P. Georgiou, "Dilated recurrent neural network for short-time prediction of glucose concentration," in *Proc. 3rd Int. Workshop Knowl. Discov. Healthcare Data, IJCAI-ECAI*, 2018, pp. 69–73.

[84] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *J. Healthcare Informat. Res.*, pp. 1–17, 2020.

[85] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.*. Berlin, Germany: Springer, 2014, pp. 446–461.

[86] C. Dalla Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 Diabetes simulator: New features," *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 26–34, 2014.

[87] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Basal glucose control in type 1 Diabetes using deep reinforcement learning: An in silico validation," *IEEE J. Biomed. Health Informat.*, pp. 1–1, 2020.
[88] T. Zhu, K. Li, L. Kuang, P. Herrero, and P. Georgiou, "An insulin bolus advisor for type 1 Diabetes using deep reinforcement learning," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5058.
[89] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *Proc. 5th KDH Workshop, ECAI* 2020, pp. 71–74.
[90] M. Porumb, S. Stranges, A. Pescapè, and L. Pecchia, "Precision medicine and artificial intelligence: A pilot study on deep learning for hypoglycemic events detection based on ECG," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, 2020.
[91] M. D. Abràmof *et al.*, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Invest. Ophthalmol. Vis.l Sci.*, vol. 57, no. 13, pp. 5200–5206, Oct. 2016.
[92] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.
[93] A. ElTanboly *et al.*, "A computer-aided diagnostic system for detecting diabetic retinopathy in optical coherence tomography images," *Med. Phys.*, vol. 44, no. 3, pp. 914–923, 2017.
[94] R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017.
[95] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.
[96] S. Keel *et al.*, "Feasibility and patient acceptability of a novel artificial intelligence-based screening model for diabetic retinopathy at endocrinology outpatient services: A pilot study," *Sci. Rep.*, vol. 8, no. 1, pp. 1–6, 2018.
[97] J. Krause *et al.*, "Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy," *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.
[98] S. Wan, Y. Liang, and Y. Zhang, "Deep convolutional neural networks for diabetic retinopathy detection by image classification," *Comput. Elect. Eng.*, vol. 72, pp. 274–282, 2018.
[99] P. Ruamviboonsuki *et al.*, "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–9, 2019.
[100] D. S. Ting *et al.*, "Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: A multi-ethnic study," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–8, 2019.
[101] F. Arcadu, F. Benmansour, A. Maunz, J. Willis, Z. Haskova, and M. Prunotto, "Deep learning algorithm predicts diabetic retinopathy progression in individual patients," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–9, 2019.
[102] I. Wittler, X. Liu, and A. Dong, "Deep learning enabled predicting modeling of mortality of diabetes mellitus patients," in *Proc. Practice and Experience Adv. Res. Comput. Rise Mach.*, 2019, pp. 1–6.
[103] T. Yamada *et al.*, "Myocardial infarction in type 2 diabetes using sodium-glucose cotransporter-2 inhibitors, dipeptidyl peptidase-4 inhibitors, or glucagon-like peptide-1 receptor agonists: Proportional hazards analysis by deep neural network-based machine learning," *Curr. Med. Res. Opin.*, pp. 1–1, 2019.
[104] I. Cruz-Vega, D. Hernandez-Contreras, H. Peregrina-Barreto, J. D. J. Rangel-Magdaleno, and J. M. Ramirez-Cortes, "Deep learning classification for diabetic foot thermograms," *Sensors*, vol. 20, no. 6, 2020, Art. no. 1762.
[105] B. M. Williams *et al.*, "An artificial intelligence-based deep learning algorithm for the diagnosis of diabetic neuropathy using corneal confocal microscopy: A development and validation study," *Diabetologia*, vol. 63, no. 2, pp. 419–430, 2020.
[106] E. W. Gregg, N. Sattar, and M. K. Ali, "The changing face of diabetes complications," *Lancet Diabetes Endocrinol.*, vol. 4, no. 6, pp. 537–547, 2016.
[107] J. W. Yau *et al.*, "Global prevalence and major risk factors of diabetic retinopathy," *Diabetes Care*, vol. 35, no. 3, pp. 556–564, 2012.
[108] Kaggle, Inc, "Diabetic retinopathy detection," 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection
[109] LaTIM Laboratory and the Messidor program partners, "Messidor-2," 2020. [Online]. Available: http://www.adcis.net/en/third-party/messidor2/
[110] PANR-TECSAN-TELEOPHTA, "E-ophtha," 2020. [Online]. Available: http://www.adcis.net/en/third-party/e-ophtha/
[111] R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, and S. Niebles, "The AI Index 2019 Annu. Rep.," 2019.
[112] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, p. 160035, 2016.
[113] University of Padova, "BioImLab," 2020. [Online]. Available: http://bioimlab.dei.unipd.it/
[114] D. Hernndez-Contreras, H. Peregrina-Barreto, J. Rangel-Magdaleno, and F. Renero-Carrillo, "Plantar thermogram database for the study of diabetic foot complications," 2019. [Online]. Available: http://dx.doi.org/10.21227/tm4t-9n15
[115] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet pre-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4918–4927.
[116] S. M. Lundberg *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 749–760, 2018.
[117] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[118] T. Kushner, S. Sankaranarayanan, and M. Breton, "Conformance verification for neural network models of glucose-insulin dynamics," in *Proc. 23rd Int. Conf. Hybrid Syst.: Comput. Control*, 2020, pp. 1–12.
[119] E. Beede *et al.*, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–12.