This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2967546, IEEE Journal of Biomedical and Health Informatics

A Deep Neural Network Application for Improved Prediction of HbA_{1c} in Type 1 Diabetes

Aleksandr Zaitcev, Mohammad R. Eissa, Zheng Hui, Tim Good, Jackie Elliott and Mohammed Benaissa, *Senior Member, IEEE*

***Abstract*— HbA_{1c} is a primary marker of long-term average blood glucose, which is an essential measure of successful control in type 1 diabetes. Previous studies have shown that HbA_{1c} estimates can be obtained from 5-12 weeks of daily blood glucose measurements. However, these methods suffer from accuracy limitations when applied to incomplete data with missing periods of measurements. The aim of this work is to overcome these limitations improving the accuracy and robustness of HbA_{1c} prediction from time series of blood glucose. A novel data-driven HbA_{1c} prediction model based on deep learning and convolutional neural networks is presented. The model focuses on the extraction of behavioral patterns from sequences of self-monitored blood glucose readings on various temporal scales. Assuming that subjects who share behavioral patterns have also similar capabilities for diabetes control and resulting HbA_{1c} , it becomes possible to infer the HbA_{1c} of subjects with incomplete data from multiple observations of similar behaviors. Trained and validated on a dataset, containing 1543 real world observation epochs from 759 subjects, the model has achieved the mean absolute error of 4.80$\pm$0.62 mmol/mol, median absolute error of 3.81$\pm$0.58 mmol/mol and R^2 of 0.71 $\pm$ 0.09 on average during the 10 fold cross validation. Automatic behavioral characterization via extraction of sequential features by the proposed convolutional neural network structure has significantly improved the accuracy of HbA_{1c} prediction compared to the existing methods.**

***Index Terms*— Convolutional neural networks, diabetes, feature extraction, machine learning, regression analysis**

I. Introduction

DIABETES mellitus represents a group of chronic metabolic disorders affecting more than 451 million people worldwide and costing more than 850 billion USD in total healthcare expenditure in 2017 [1]. Diabetes is associated with increased levels of blood glucose (BG) and in type 1 diabetes (T1D) there is an absolute deficiency in insulin production, which has to be substituted by multiple daily injections of insulin accompanied by regular Self-Monitored Blood Glucose (SMBG) measurements or alternatively Continuous Glucose Monitoring (CGM). The goal of such a regimen is to keep BG levels within a near normal range (usually between 4 and 10 mmol/L), therefore avoiding both the consequences of high glucose levels (hyperglycaemia) and the dangers of low BG (hypoglycemia). Since the early 1980s the general success of glycaemic control has been clinically assessed using a laboratory measure of glycosylated hemoglobin (HbA_{1c}), which reflects the long-term average blood glucose levels [2]. It has been clinically shown that lower HbA_{1c} levels (below 59 mmol/mol) are associated with significantly smaller risks of serious diabetes-related complications such as retinopathy (eye damage) nephropathy (kidney damage) and nerve damage (neuropathy) [3]. Considering that HbA_{1c} measurement is usually carried out once every three to six months, there is a clinically justified motivation to provide diabetes patients and their clinicians with an ongoing HbA_{1c} estimate in order to reflect the quality of their glycaemic control and reduce the associated risks by facilitating prompt behavioral interventions if needed.

SMBG measurements, typically taken 3-8 times a day, provide an insight into the quality of patient glycaemic control, which indicates the resulting HbA_{1c} . A number of studies have identified the links between BG measurements over a period of 5-12 weeks and a subject's resulting HbA_{1c} [4], [5], [6], [7], [8]. Analysis of data from the Diabetes Control and Complications Trial [6] allowed identification of a strong linear relationship between the average measured blood glucose and HbA_{1c} . The work of D.M. Nathan, *et al.* [7] has confirmed these findings and defined a mathematical relationship between the mean average of SMBG readings and glycosylated hemoglobin, which allows for a simple prediction of HbA_{1c} given sufficient amount of BG measurements per day. More recently B.P. Kovachev, *et al.* [8], [9] proposed a more sophisticated non-linear HbA_{1c} prediction model based on the dynamic tracking of daily SMBG profiles. However, the aforementioned prediction methods have been developed and validated on a highly selective sample of patients, excluding those with poor glycaemic control and complications, while also often requiring the subjects to conform to restrictive measurement and calibration protocols. Besides, biological variations may also undermine the HbA_{1c} estimation [10]. As a result, these models suffer from significant prediction inaccuracy when applied to imperfect SMBG data which can be often observed in patients outside experimental settings.

Manuscript received January 15, 2020. This project is partly funded by the National Institute for Health Research (NIHR), [DAFNEplus (RP-PG-0514-20013)]. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. *(Aleksandr Zaitcev and Mohammad R. Eissa are co-first authors.)(Corresponding author: Aleksandr Zaitcev.)*

A. Zaitcev, M. Eissa, Z. Hui, T. Good and M. Benaissa are with the Department of Electronic and Electrical Engineering, University of Sheffield, S1 4DE, UK (emails: a.zaitcev@sheffield.ac.uk; m.eissa@sheffield.ac.uk; zhui2@sheffield.ac.uk; tim.good@sheffield.ac.uk; m.benaissa@sheffield.ac.uk)

J. Elliott is with the Department of Oncology and Metabolism, University of Sheffield, S10 2RX, UK (e-mail: j.elliott@sheffield.ac.uk)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2967546, IEEE Journal of Biomedical and Health Informatics

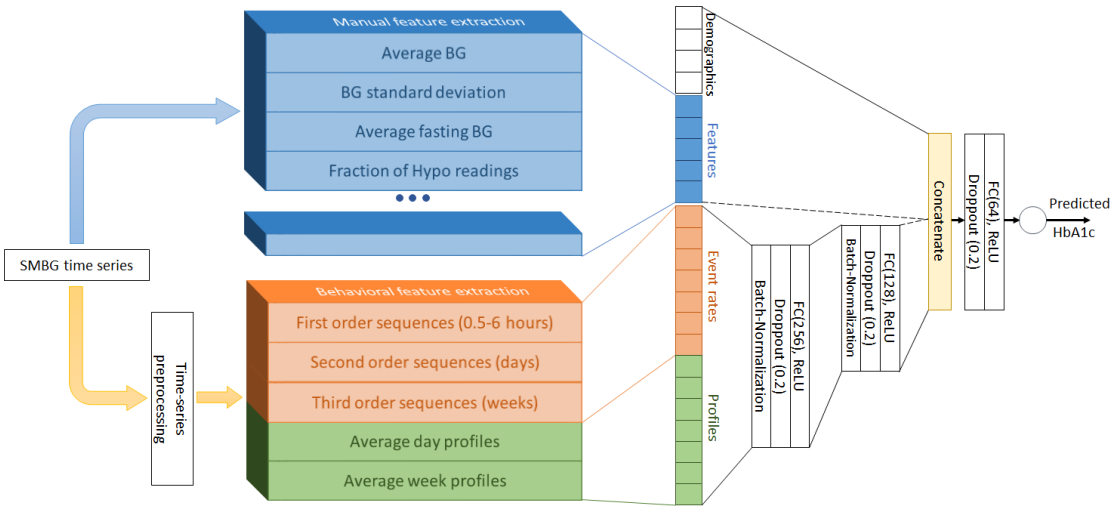

Fig. 1. Top level architecture diagram of the model.

This work presents a novel HbA_{1c} estimation model in T1D with a particular focus on practical applicability and robustness in the presence of realistically imperfect data and outliers. This approach is based on the assumption that a subject's physiological specificities, behavior and various aspects of glycaemic control are holistic in nature, i.e, can be viewed as an interconnected system. The times of day and week when SMBG measurements are taken, specific sequences of BG levels, and regularities of such sequences provide an insight into the subject's habits, life schedule and the overall behavior. By also assuming that patients who share common behavioral patterns are likely to have similar quality of glycaemic control and perhaps even physiological features, the comprehensive characterization of behaviors observable in data may provide factors supplementary to the average BG that help regularizing the HbA_{1c} prediction and thus achieve better precision.

Such complex analyses have become feasible with the recent advances in artificial neural networks and deep learning, which has provided research communities with new tools for reliable data-driven modelling of non-linear predictor relationships and their automatic extraction from the available observations [11]. Our proposed model is a feed-forward network which utilizes novel architectures of convolutional neural networks (CNNs) and fully connected (FC) layers to perform a combined analysis on SMBG time series alongside the conventional manually derived metrics of glycaemic control.

II. The Model

The proposed approach treats the T1D HbA_{1c} prediction task as a data-driven regression problem that is solved using a hybrid model that takes input in the form of SMBG epochs with an addition of relevant patient demographics data such as age, gender and the duration of diabetes (Figure 1). The model was implemented as a Tensorflow [12], [13] computational graph consisting of two major branches: one being a CNN implementing behavioral feature extraction from time series and one for extraction of known correlates of HbA_{1c} . The outputs from these two branches are concatenated together with demographics features and passed to the output structure consisting of fully connected (FC) layers, which perform feature fusion and produce the regression output. The following sections provide details about the dataset, preprocessing procedures, and discuss the manual feature selection and neural network architectures employed in the proposed model.

A. Time series preprocessing

For people with T1D it is recommended to test blood glucose levels at least 4 times per day - before each meal and once before bed. Besides that, they are encouraged to take additional blood tests before, during and after exercise, before driving, when symptoms of hypoglycaemia are present, before high-risk activities and during periods of illness [14]. In practice people with T1D do not always follow such recommendations and as a result the sampling rate of SMBG series ranges from 2 to 8 measurements per day on average. The proposed time series preprocessing aims to highlight the timing structure of data, convey semantics of individual readings, while also preserving its sequential nature and allowing for cross-subject training. This was achieved by transforming the original irregular time series into an interpolated fuzzy sparse format.

In order to align the training epochs and standardize their lengths, SMBG data was first interpolated onto a regular time grid with 30 minute resolution using the piecewise cubic Hermite interpolating polynomial (PCHIP) [15]. Individual time series were cut and zero-padded in such a way so that all epochs ended at 3:00 am on Monday nearest to the HbA_{1c} test and started exactly 84 days before that. An epoch length of 12 weeks was chosen based on the previous studies on HbA_{1c} prognosis [7].

Next, a certainty mask was calculated as 1D time series with a value of 1 at times of BG measurements and a value of 0 for samples further than 1.5 hours from measurements (Figure 2-A). Certainty levels below 0.90 were set to zero to produce a sparse mask that was piecewise multiplied with the smoothed SMBG to obtain an interpolated sparse SMBG representation (Figure 2-B).

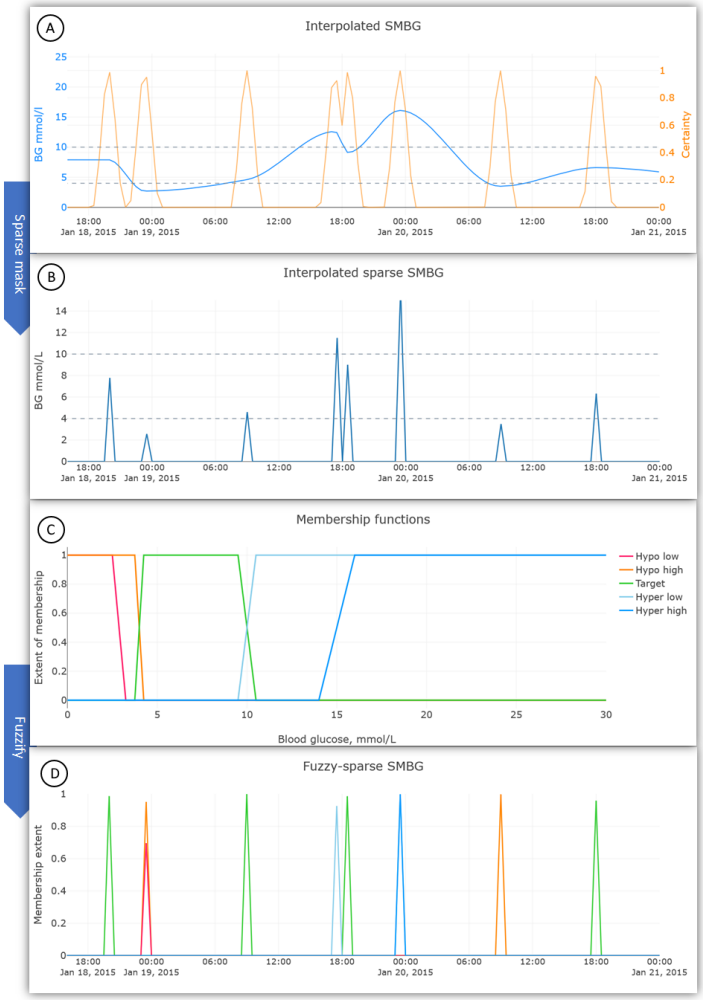

Fig. 2. Schematic diagram of time series preprocessing stages.

The measured blood glucose levels that constitute these time series typically range between 1.5 to 30 mmol/L, with values below 4.0 mmol/L meaning hypoglycemia and values above 10.0 mmol/L meaning increased glucose toxicity or hyperglycemia. Different levels of blood glucose result in different perceived symptoms and hence, when measured by subjects, induce dissimilar reactions. In order to accentuate this, the interpolated sparse SMBG series were allocated to different classes according to their values, or in other words converted to fuzzy variables. The ranges for fuzzy classes were selected according to the thresholds used in clinical practice with BG values below 3.0 mmol/L meaning clinically significant hypoglycemia, below 4.0 mmol/L - an alert for hypoglycemia, 4.0-10.0 mmol/L - target range, above 10.0 mmol/L - general hyperglycemia and above 15.0 mmol/L - severe hyperglycemia (Figure 2-C). As a result, the time series were transformed into the desired ANN input format with: values normalized to $[0, 1]$; preserved timing and sequential structure; time grid aligned between observations; and highlighted semantics of various BG ranges (Figure 2-D). Thus, the input for the CNN branch was represented by a 3D tensor $\mathbf{D} \in \mathbb{R}^{N_e \times T \times N_f}$, where N_e is the number of available epochs, T is their length and N_f is the number of fuzzy classes employed. In our settings $\mathbf{D} \in \mathbb{R}^{1543 \times 4032 \times 5}$.

B. Network architecture elements

The following sections give a brief overview of the proposed architecture's elements.

1) Feed-forward neural network: Feed-forward neural networks (FNNs), an example of which is the proposed model, are essential types of structures in deep learning. For a regression problem with input $\boldsymbol{x}$ and a regression target y the aim is to find a mapping function f^*, so that $y = f^*(\boldsymbol{x})$. A feed-forward network defines an approximation $\hat{y} = f(\boldsymbol{x}, \boldsymbol{\theta})$, where parameters $\boldsymbol{\theta}$ are learned from the available observations of $\boldsymbol{x}$ and y [16]. The mapping f is generally non-linear and is implemented by series of many various functions $f^{(1)}, f^{(2)}, ..., f^{(M)}$. The structure of FNN can be described as an acyclic computational graph, where information flows in one direction from input $\boldsymbol{x}$ through a network of nodes implementing mapping f to the output $\hat{y}$, hence the term **feed-forward** in FNN.

2) Fully connected layer: Fully connected (FC) or Dense layers [16] multiply the input vector with a weight matrix or kernel $\boldsymbol{W}$ and add a bias term b, which is often followed by the application of a non-linear activation function. Both $\boldsymbol{W}$ and b are learned from data during the neural network training. In our setting the output at layer l is defined by:

$$\boldsymbol{x}^l = \mathrm{ReLU}(\boldsymbol{W}\boldsymbol{x}^{l-1} + b) \,, \tag{1}$$

where ReLU is the Rectified Linear Unit activation which in effect cuts off negative elements of the input [17]:

$$\mathrm{ReLU}(\boldsymbol{x}) = max(0, \boldsymbol{x}) \,. \tag{2}$$

3) Convolution layer: Convolutional layers extract local information from the regularly sampled data such as time series, images or video samples [16]. Each convolutional layer consists of a number of filters which are convolved with the input to produce activation maps describing the extent of features' presence in data. The convolution is usually followed by the application of bias and non-linear activation. The filter size, also called receptive field, is usually small compared to the size of the input and fixed, defining the scale of features extracted. In general terms the output of the j-th filter at 1D convolutional (Conv1D) layer l is obtained as:

$$\boldsymbol{x}_j^l = \mathrm{ReLU}(\sum_{i \in D} \boldsymbol{x}_i^{l-1} * \boldsymbol{w}_{ij}^l + b_j^l) \,, \tag{3}$$

where D is the depth of input (number of channels), $\boldsymbol{w}_{ij}^l$ is the j-th filter kernel for i-th input dimension and b_j^l is the corresponding bias term. Similarly this definition can be extended to higher dimensional input to implement 2D convolutional (Conv2D) or 3D convolutional (Conv3D) layers.

Convolutional filter kernels encode the types of features being extracted. By stacking multiple convolutional layers it is possible to learn features at various levels of abstraction. In order to support that, convolutional nodes are commonly alternated with pooling layers, which downsample the data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2967546, IEEE Journal of Biomedical and Health Informatics

thus increasing the next convolution's receptive field [16]. Pooling operations subsample input and reduce the extracted subsets. For example, with a 1D input of length T $\mathbf{d} \in \mathbb{R}^T$ a 1D max pooling operation with stride s, pool size p and zero-padding to length $\hat{T}$ can be defined as follows:

$$\begin{aligned} \mathbf{y} &= \text{maxpool}(\mathbf{d}) \in \mathbb{R}^{\frac{\hat{T}}{s}} \\ \mathbf{y}_i &= \max(\mathbf{d}_i)\,, \mathbf{d}_i \in \mathbb{R}^p\,, \end{aligned} \tag{4}$$

where $\mathbf{d}_i$ is the i-th slice of size p.

Features extracted by 1D CNN at various scales are still represented by time series, meaning that data is high dimensional, sequential and features from different epochs may often have different offsets. The fully connected layers that process CNN features and produce the regression output are highly susceptible to such effects making the direct analysis of temporal data problematic. This issue can be resolved by introducing a global average pooling operation at the interface between CNN and FC layers which quantifies the average extent of feature's presence over time. In order to improve robustness in the presence of missing data the proposed model employs a variation of this operation: a weighted global average pooling (WGAP), which uses the certainty mask from interpolation as weights for averaging. With 1D input of length T $\mathbf{d} \in \mathbb{R}^T$, and $\boldsymbol{w}$ holding values of the corresponding certainty mask, the scalar representing the weighted average is given as:

$$\mathbf{y} = \text{WGAP}(\mathbf{d}) = \frac{\sum\limits_{t \in T} d_t w_t}{\sum\limits_{t \in T} w_t}\,. \tag{5}$$

4) Regularization: In machine learning regularization is applied in order to prevent overfitting, which occurs when a model becomes highly specialized to training observations instead of learning the generative model of data. A common way to regularize a neural network is to apply a Gaussian (L2), Laplacian (L1) or their combined penalty (Elastic Net [18]) to the learned kernels at different layers. For a cost function $J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y})$ with network parameters $\boldsymbol{\theta}$, training set $\boldsymbol{X}$ and targets $\boldsymbol{y}$ the application of Elastic Net regularization to the j-th layer updates the cost function into:

$$\hat{J}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) = J(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \|\boldsymbol{w}\|_2\,, \tag{6}$$

where λ_1 and λ_2 are the hyperparameters defining the extent of regularization. With λ_1 being large enough the regularized kernels become sparse, which is useful for dimensionality reduction and feature selection tasks [16].

Besides the norm penalties, another common regularization mechanism applied in deep neural networks is **dropout**. Dropout is a technique recently proposed by Srivastava, et al. [19], where a number of neurons are randomly ignored during training temporarily removing their contribution to information flow and weight updates. This reduces the extent of neuron co-adaptation enforcing the learning of data generative model rather than specialization to specific values, which in effect leads to a better model generalization.

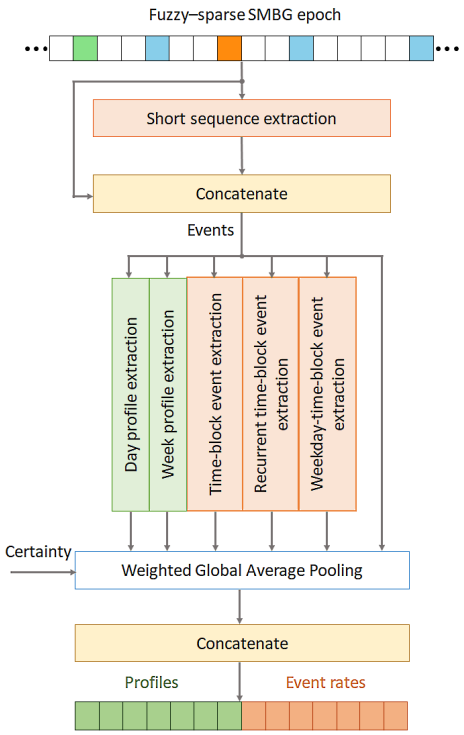

Fig. 3. Architecture of behavioral pattern extraction network.

C. Behavioral feature extraction

The times of BG measurements, the bands of values, their specific sequences and recurrence patterns may carry information about the various problems with BG regulation while also characterizing lifestyles and schedules of subjects. In practice clinicians often visually analyze such patterns in patients' SMBG diaries in order to identify certain difficulties with BG control and provide advice or behavioral intervention [20]. For instance, a high BG reading followed by a hypoglycemic event within 4 hours generally means incorrectly selected bolus (quick acting) insulin dosage in an attempt to correct hyperglycemia; low BG level in the morning followed by in-target glucose before bed can often mean a mistake in basal (long acting) insulin dose selection; recurring hypoglycemia on Saturday mornings may often point out a BG regulation problem associated with alcohol intake, etc. The hypothesis in this paper, is that the rates of such sequential event occurrences and their recurrence patterns characterize certain problems with BG control and when combined represent the overall quality of a subject's glycaemic control and their behavior.

In this approach, behaviors are quantified using the rates of SMBG events that can be either separate BG tests or their short sequences. Besides the basic rates of occurrence, specific times of day and days of week when certain sequences take place must also be highlighted if such notions are significant for the resulting HbA_{1c} value. Such feature extraction is supported by a set of multiscale convolutional layers and also daily and weekly average BG profiles. The diagram in Figure 3 shows the architecture of the proposed behavioral feature extraction network that will be discussed in the following subsections.

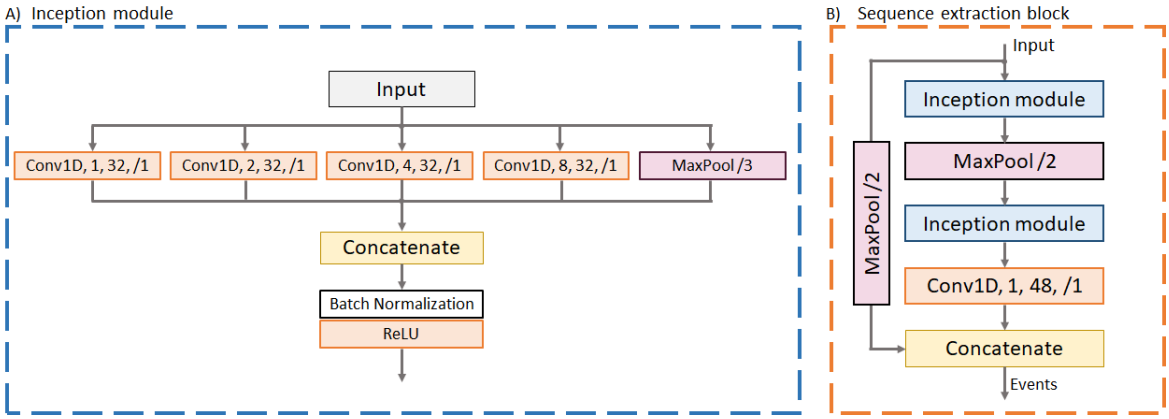

Fig. 4. Architectures of inception module and event extraction sequence.

1) Short sequence extraction: One dimensional convolutional layers (Conv1D) have the capability to extract local features from time series at a single time scale determined by the fixed filter size. By manually setting the convolutional kernel dimensions it is possible to explicitly define the time scales of locally extracted features. Such quality of temporal convolutional layers is the key reason why they were chosen as a basis for multiscale feature extraction instead of recurrent neural network layers [21], that are also commonly used for time series modelling, but lack the flexibility in time scale configuration.

The characteristic short sequences of SMBG readings generally take place within 0.5 to 6 hours. Considering such varying time scales, extraction of these sequences by the means of a single Conv1D layer is problematic. To address this issue, we have utilized an architecture similar to the GoogLeNet Inception module [22] that included a number of parallel Conv1D filters of different scales. Figure 4 shows the diagram of our proposed variation of Inception module (A) and how it was employed in the event extraction sequence (B). Here notation [Conv1D, L, N, /S] represents a 1D convolutional layer with filter size L, filter number N applied with strides (time shift step) S, and [MaxPool /S] stands for a Max Pooling operation with pool size S, which in effect downsamples the input by the factor of S. The outputs of parallel operations are concatenated across the depth dimension, which is followed by batch normalization [23] and the application of a non-linear Rectified Linear Unit [17].

Our variant of Inception module consists of four parallel logarithmically scaled Conv1D filters and a MaxPool/3 operation (Figure 4-A). Here, Conv1D layer with filter size 1 represents a trainable set of linear combinations between the input channels. Considering the input sampling rate of 30 min, a sequence of two Inception modules with a downsampling operation between them creates a set of pathways, where local features of time scales between 0.5 and 8 hours can be extracted. In order to control dimensionality after the multi-scale convolutional blocks, an additional Conv1D layer with ReLU activation was employed to cut out the unwanted components and reduce the depth to 48 channels. Finally 53 types of SMBG events were formed by combining the extracted multi-scale sequences with the original fuzzy sparse time series that were downsampled in order to conform with the resulting epoch lengths. Following this processing stage each SMBG epoch is represented by a matrix of 2016×53 elements.

2) Day scale feature extraction: The time of day when an SMBG point was taken attaches additional context to the reading or sequence in which it was involved. In order to include such notion of data in the analysis, a daily BG profile estimator was used alongside a set of convolutional filters that produce features separable by time of day (Figure 5). The time block analysis pipeline starts from a downsampling operation which results in each out of 84 days of epoch being represented by 8 time blocks. This operation is necessary in order to reduce the offset between different patients' day schedules and thus allow for a cross-subject analysis. The resulting 672×53 matrices representing epochs were then reshaped into $84 \times 8 \times 53$ tensors and passed to the three different branches of time block feature extraction (Figure 5 A,B,C).

A number of studies have identified a strong relationship between the subject's hourly BG profiles and their HbA_{1c} [24], [8]. The 8-point average day profiles were obtained by averaging the aligned tensors along the days axis as given in Figure 5-A. A Conv2D layer with 16 filters of size 1×1 was used here to control the dimensionality of extracted profiles. Besides, in order to reduce the effects of days with low number or no measurements, the BG profiles were obtained as a weighted average with weights being the average certainty from each individual day of epoch (Figure 2-A). As a result, days with skipped readings and lower average of certainty mask had contributed less to the estimated BG profiles. Following the weighted averaging each profile of size 8×16 was flattened yielding 128 features.

Time block specific features such as a rate of under-corrected hypers after breakfast or a rate of nighttime hypoglycemic events may often be indicative of certain BG regulation problems. In our model such features were extracted using a 2D convolutional layer with a receptive field spanning the whole day duration (Figure 5-B). Thus 48 Conv2D filters of size 8×1 were applied to the input tensor with stride 1, i.e.,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2967546, IEEE Journal of Biomedical and Health Informatics

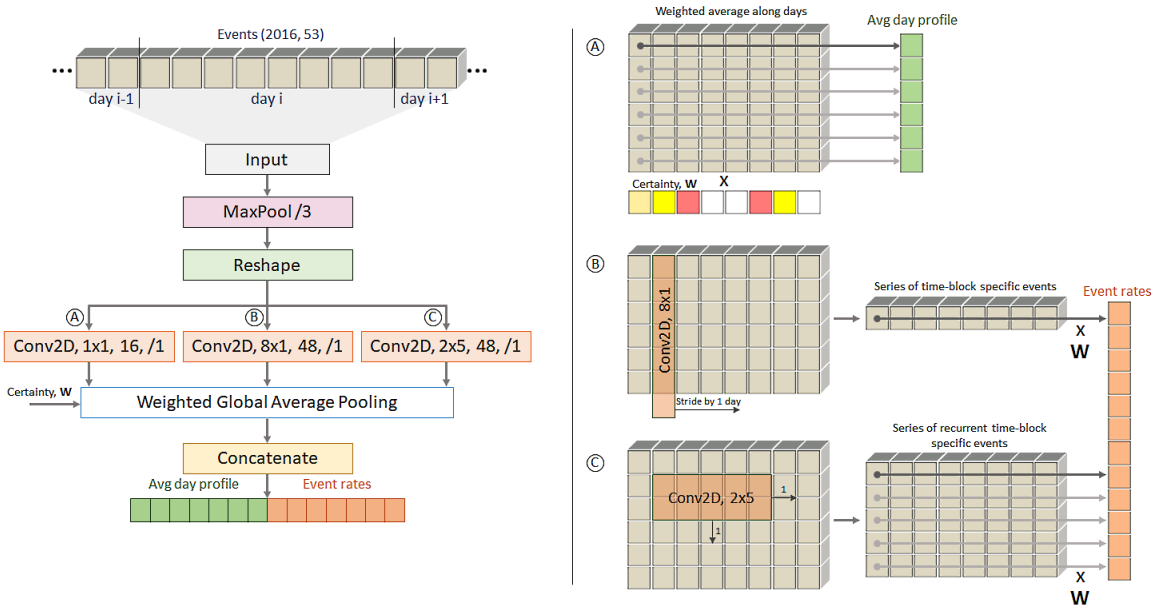

Fig. 5. Architecture and visualization of pattern extraction on a scale of days.

to each separate day, producing output in the form of 1D time-series of time-allocated events. Same as with the BG profile calculation, the rates of these events were then obtained by weighted averaging along days of epoch yielding 48 additional predictors.

Besides the global rates of time-specific events over 12 weeks of SMBG data, recurrent patterns of event occurrences are often used to characterize certain problems in patient's BG control, as for example, described in [25]. In the proposed processing pipeline such features were extracted using another Conv2D layer the receptive field of which spanned over 5 days and 2 adjacent time-blocks. As shown in the Figure 5-C, 48 filters of size 2×5 were applied to the input tensor with stride 1 along both dimensions. The time dimension of the output was then once again reduced by weighted average pooling along the day axis resulting in additional 371 features describing repetitions of time specific events in each processed SMBG epoch.

3) Week scale feature extraction: The daily life schedule of people with T1DM may also vary with respect to the days of week, e.g., weekend BG profiles may often differ from those recorded on weekdays. In the proposed model such specificities were captured using an additional network operating on the scales of weeks. As shown in Figure 6 the network extracts weekly SMBG profiles and weekday-time specific event rates similarly to the aforementioned day scale feature extraction block.

Same as before, the input of 2016×53 event time series was downsampled in a way that each day was represented by 8 time blocks. Next, event sequences were reshaped into series of weekly tensors shaped $12\times7\times8\times53$, meaning that each of the 12 weeks was represented by a $7 \times 8 \times 53$ tensor. Similar to day profiles, the average week profiles were extracted by the sequence of a convolutional layer for dimensionality reduction followed by the global average pooling along the week dimension, which resulted in 448 new features appended to the output vector (Figure 6-A). SMBG events allocated to both weekdays and time blocks were extracted using a Conv3D layer with 48 filters of size $8 \times 7 \times 1$ applied with a step time of 1 week (Figure 6-B). In the same way as before, rates of such events were obtained by global average pooling along the week dimension producing 48 additional variables.

D. Manual feature extraction

The CNN captures the underlying patterns of the time-series SMBG data. The clinical practice utilizes metrics that provide evaluation means for patients' daily regulation of BG levels. These features were extracted from 12 long week epochs and analyzed to determine their effect on HbA_{1c} levels. The metrics manually derived from data included: average BG, median BG, inter-quartile BG, the standard deviation of BG, the standard deviation of the mean of the BG, percentages of hypoglycemia (clinically significant and alert), and hyperglycemia (significant and general) [26]. Considering a patient could contribute to more than one HbA_{1c} result and our interest is to capture the behavior of each individual, a generalized estimating equation (GEE) was used to account for the correlation in the data by the subjects of the study. We ran a hierarchical analysis of the covariates and used the p-value of 0.05 to determine the significant predictors of the outcome (based on the statistics on α error). While controlling for the individual's characteristics (i.e. age, sex and diabetes duration), the average BG, the standard deviation of the mean of BG, and alert hypoglycemia and general hyperglycemia were the most significant predictors of the HbA_{1c} level. Therefore, these features were the input to a branch of the neural network.

E. Network output

The output network of the proposed model consists of multiple FC layers and regularization operations that support

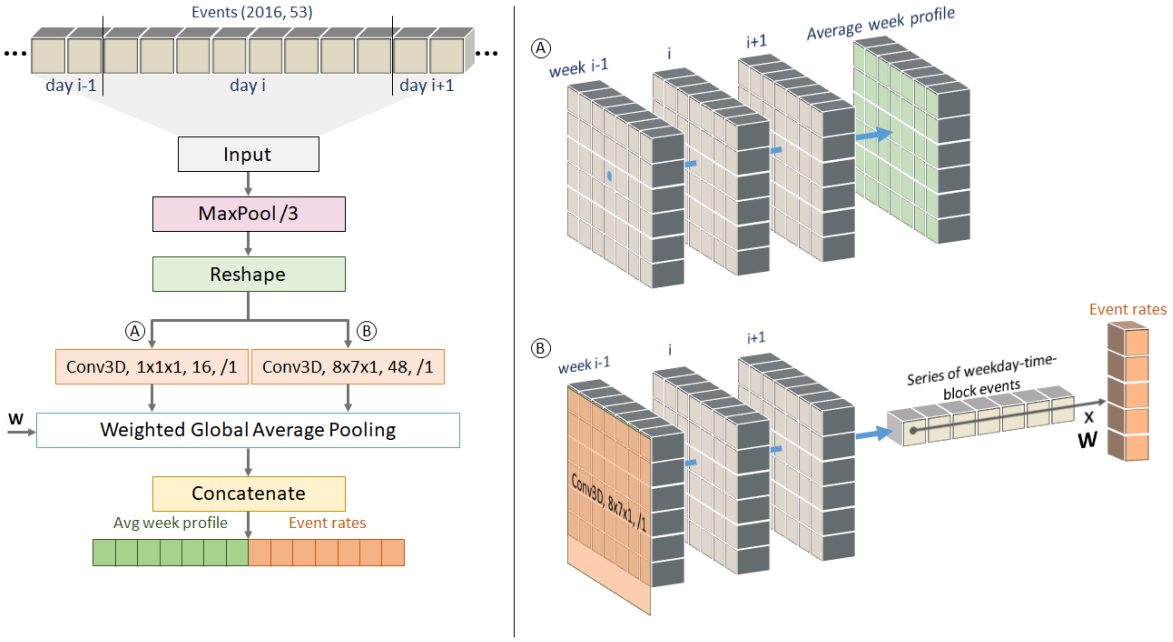

Fig. 6. Architecture and visualization of the week scale pattern extraction scheme.

dimensionality reduction and calculation of regression output (Figure 1). Its structure was selected heuristically and hyperparameters were optimized in multiple stages of grid search. A small subset of 128 out of the total 3912 CNN features is selected here using the two dropout-regularized FC layers with restricted number of units. To support this, sparse regularization penalty with $\lambda_{l1} = 0.01$ was also applied to the weights of FC kernels [27]. Besides that, Batch Normalization [23] was also applied in the output FC layers, as shown in Figure 1. Same as in [28], the motivation behind it was to increase the convergence speed and stability during the model training. Next, the reduced CNN features were concatenated with manually extracted predictors and then a combined processing was performed by a single dense layer with 64 units followed by a standard linear regression output unit.

III. Experiment and Results

A. Data set

The proposed model was trained and validated on a dataset collected from 759 people with T1D attending Sheffield Teaching Hospitals in the period between 2013 and 2015. Each entry in the training set was formed by a subject's HbA_{1c} test, their demographic data (age, gender, years with diabetes) and final 84 days (12 weeks) of SMBG readings preceding the test, with the length of epochs being justified by physical meaning and definition of HbA_{1c} [2]. In practical settings SMBG data provided by patients is often incomplete, containing days with skipped readings or even intervals of days with no BG measurements at all. In order to validate the practicality of our method we have allowed such imperfect epochs to be included in our dataset. Thus, SMBG time series containing no more than 30 % of days with completely missing data and at least 2.5 BG tests per day on average on days with measurements were included, which resulted in a total of 1543 epochs being selected and 351 epochs being filtered out.

Selected epochs had a mean of 3.67 readings per day and 89% of days with any recordings. The average age of the 759 selected participants was 48.76 ± 17.2 with a mean duration of diabetes of 24.65 ± 15.57 years. Among them 392 were male and 367 were female ($51/49\%$). At least 29.2% of participants have attended a DAFNE course on T1D management [29]. HbA_{1c} levels were measured in a laboratory using SEBIA CAPILLARYS kit, with error coefficient of variance (CV) of $1.8 - 2.4\%$ [30]. Selected HbA_{1c} test results had the mean of 66.84 ± 11.36 mmol/mol.

B. Accuracy analysis and comparison

The proposed model was trained and evaluated iteratively on the selected dataset. The evaluation protocol involved ten-fold cross validation (CV), where the dataset was randomly split multiple times so that each data point was used in a training set 9 times and once in a test set. A more comprehensive evaluation would be possible using, for example, "Leave One Group Out" partitioning approach, where each single patient's data would be used as a validation set in a separate train-evaluation scenario. However, such exhaustive validation approach is extremely computationally expensive and is infeasible in our settings. The further validation with external data is also of interest.

Within each fold of CV, the proposed model was trained using the Adam optimization algorithm [31] for 250 epochs using the batch size of 96 observations. The training and validation learning curves describing the proposed model training progression are displayed in Figure 8. Note that the non-smooth optimization trajectory is due to the effect of the Dropout layers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2020.2967546, IEEE Journal of Biomedical and Health Informatics

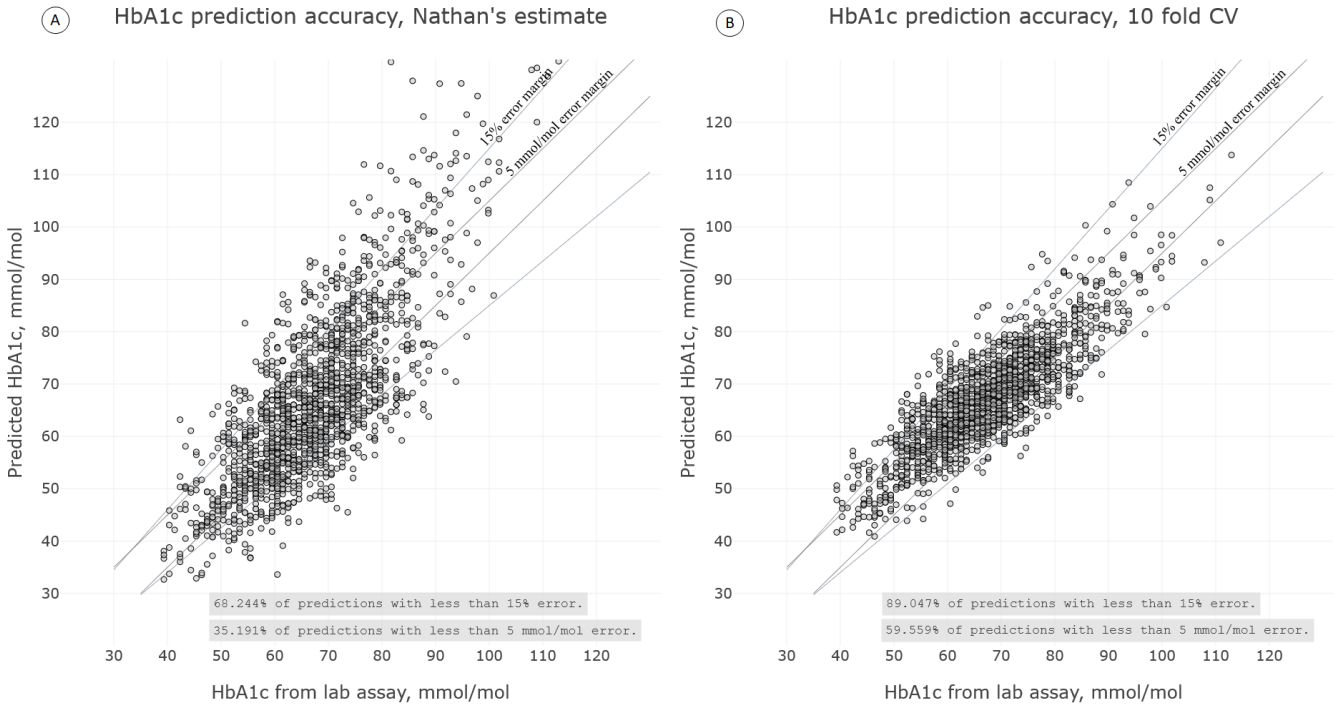

Fig. 7. A)-B) Error grid for Nathan's HbA_{1c} estimate and the 10 fold CV of the proposed model.

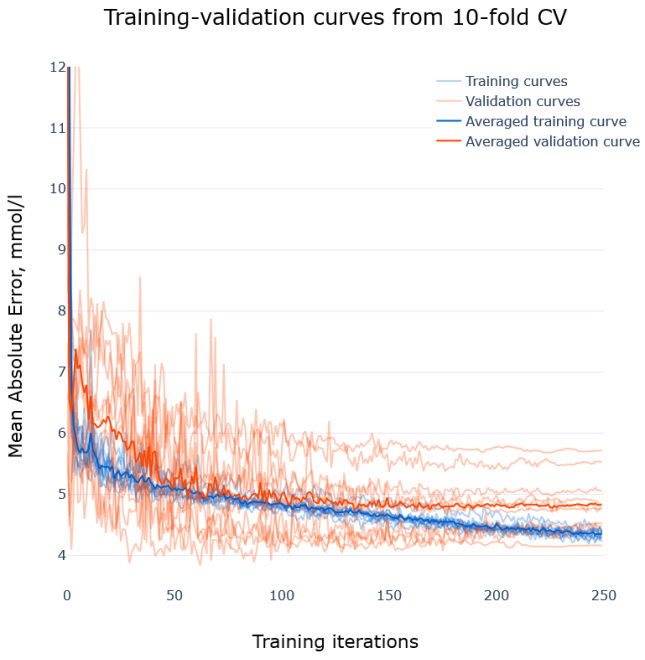

Fig. 8. Summary of the proposed model training progression from 10 fold CV. Training and validation MAE loss learning curves are given in blue and orange respectively.

During the 10-fold CV partitioning the observations were grouped by patient ids to avoid the effects of individual specificity. Besides the whole model, the manual features and the CNN branches of model were also evaluated separately for comparison. Accuracy metrics collected over the validation folds are summarized in table I. These results averaged over the ten folds are also compared to the accuracy of the HbA_{1c} estimation formula described by DM Nathan, *et al.* [7]. Comparison with the dynamic tracking method of B.P. Kovachev, *et al.* [8], [9] is of interest, but are not provided here, since the method requires 7-point BG profiles and regular calibration points which are not available in the given dataset. In this table MAE stands for mean absolute error, MedAE denotes median absolute error, bias - average prediction bias and R^2 is the coefficient of determination.

Over the 10 validation cycles our model has achieved the average MAE of 4.80 ± 0.62 mmol/mol with a median at 3.81 ± 0.58 mmol/mol, R^2 of 0.71 ± 0.09 and a total regression bias of -0.0071 ± 0.21. Figure 7-A and Figure 7-B show the error grids of predictions accumulated over the 10 folds compared to the estimates obtained with Nathan's formula. With the proposed model prediction percentage error for 89 % of observations did not exceed 15 % while 59.5 % of estimates fell within a 5 mmol/mol error margin.

Table II gives precision metrics of the proposed model and linear estimator by ranges of reference HbA_{1c} . Within all presented bands our model has shown an improvement in

TABLE I
VALIDATION ACCURACY COMPARISON

	Nathan's formula	Manual FE network	CNN only	Combined
MAE	8.26	6.10 ± 0.84	5.98 ± 0.81	4.80 ± 0.62
MedAE	7.08	5.03 ± 0.66	4.87 ± 0.64	3.81 ± 0.58
R^2	0.19	0.58 ± 0.07	0.62 ± 0.07	0.71 ± 0.09
Bias	-0.43	-0.18 ± 0.2	-0.19 ± 0.34	-0.0071 ± 0.21

Accuracy metrics: MAE = mean absolute error, MedAE = median absolute error, R^2 = coefficient of determination. Standard deviation between the folds of cross validation is provided where appropriate.

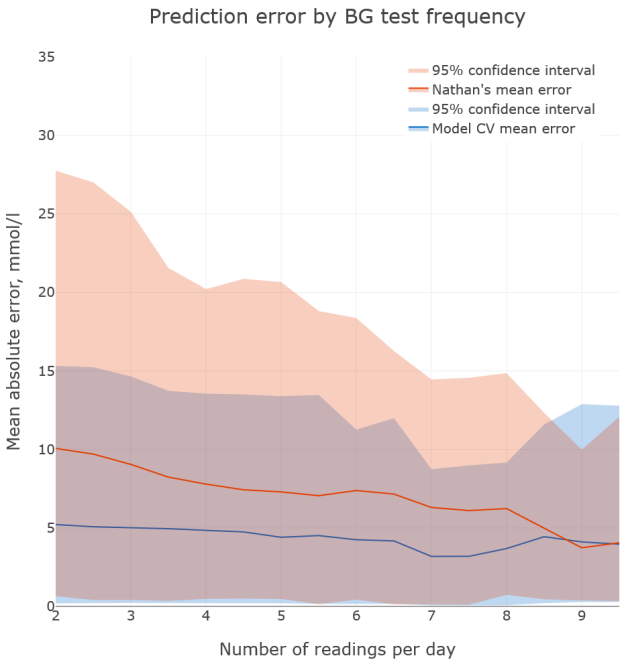

Fig. 9. Mean absolute prediction errors and their 95% confidence intervals of the proposed model (in blue) and Nathan's estimate (in orange) plotted against the average BG test frequency.

TABLE II
VALIDATION ACCURACY OF COMBINED MODEL BY HbA_{1c} RANGES

The proposed combined model			
HbA_{1c} ranges, mmol/mol	≤ 58	58 - 87	≥ 87
Count	300	1166	77
MAE	5.09 $\pm$0.78	4.58 $\pm$0.55	6.94 $\pm$1.71
MedAE	3.93 $\pm$0.48	3.60 $\pm$0.49	6.51 $\pm$1.1
Bias	3.85 $\pm$1.62	-0.66 $\pm$0.87	-5.11 $\pm$3.19
Nathan's formula			
MAE	6.32	8.39	13.92
MedAE	5.35	7.26	13.11
Bias	-0.35	-0.91	6.58

Accuracy metrics: MAE = mean absolute error, MedAE = median absolute error. Standard deviation between the folds of cross validation is provided where appropriate.

prediction accuracy and most notably in the 58-87 mmol/mol range. From this table it can also be noted that the generalization accuracy of the proposed model in a particular band of HbA_{1c} highly depends on the amount of relevant observations.

Figure 9 displays the 95% confidence intervals of prediction errors accumulated over the 10 validation folds against the average amount of SMBG readings per day. The error means and 95% confidence intervals for each method are given in blue for the proposed model and in orange for Nathan's estimate. The accuracy comparison shows that the proposed model consistently outperformed the linear Nathan's estimate on the given dataset, and especially on sparse epochs with lower density of measurements. With higher density of SMBG the information about the true average level of BG becomes clearer, thus linear HbA_{1c} estimation on epochs with dense measurements provides accuracy comparable to our model.

IV. DISCUSSION

The modelling approach presented in this work has shown a high HbA_{1c} prediction accuracy and robustness in the presence of incomplete data. Through a hybrid feature extraction mechanism SMBG series preceding the HbA_{1c} tests were encoded to reflect the long term BG dynamics and subjects' behaviors. The latter was represented by a combination of BG profiles and sequential event rates extracted using a set of trainable time domain filters. Considering that the average BG alone is sufficient for a relatively accurate HbA_{1c} estimation, it can be stated that the increase in precision was achieved by the additional regularization and biasing performed by the output FC layers based on the shared subject behaviors. Similar to many other data-driven approaches the model's performance is expected to improve further with the increase in size of the training sample.

Although the dataset used in this work is relatively large and more inclusive compared to the other related works, the number of consecutive HbA_{1c} tests for individual subjects was limited. Some of the known frameworks, such as eA1c described in [8], use periodic laboratory tests for calibration and then the relative change in HbA_{1c} is predicted rather than its absolute value. Thus in further work this implies that the use of such calibration points would allow for a more personalized behavioral pattern analysis, which would be expected to improve the model stability and potentially even further reduce the data integrity constraints.

The availability of accurate daily HbA_{1c} estimates and their short-term trends can allow for more prompt and personalized behavioral interventions and advice from clinicians helping to shape the lifestyles of patients, reducing the risks of diabetes and improving their quality of life. Specific behavioral patterns that lead to poor glycaemic control and increase in HbA_{1c} can be extracted for example from the CNN branch outputs, although due to the noisy nature of data, high model capacity and perhaps insufficient amount of observations, the visual assessment of such patterns is problematic. Therefore, further analysis of behavioral features extracted in CNN is of interest.

For the majority of people with T1D a large number of daily BG measurements allows for a better accuracy of HbA_{1c} prediction even with simpler methods such as linear regression since a more complete picture about the average BG levels is provided [32]. With CGM equipment becoming increasingly available in clinical practice the accurate estimation of HbA_{1c} based on average blood glucose becomes feasible. A number of studies, however, have identified that average BG alone may sometimes be insufficient for accurate prediction of glycosylated hemoglobin due to the certain physiological specificities of subjects [10], [33]. We speculate that such biological traits affect the BG patterns and patient behaviors at various scales and therefore, were captured by our feature extraction mechanism and included in the automatic analysis by FC layers resulting in reduced prediction error. Considering the nature of CNN, our proposed modeling method can seamlessly be adapted to CGM data and other signal modalities allowing for even more accurate daily tracking of HbA_{1c} and other novel applications of behavioral pattern analysis.

V. Conclusion

The dynamic tracking of HbA_{1c} supports more timely and informed behavioral interventions in cases of poor self-care in T1D. This study has presented a hybrid deep learning model for HbA_{1c} prediction from realistically imperfect time series of SMBG. While major indicators of HbA_{1c} are well-known and can be easily obtained, their predictive performance deteriorates in the presence of incomplete data and certain physiological specificities. The automatic extraction of self-care behavior patterns by the proposed multi-scale temporal CNN provides additional context for regression and regularizes the prediction based on behavioral clustering.

The results of regression analysis on a sample of 1543 observations from 759 subjects with T1D has demonstrated the advantages of behavioral characterization. The accuracy of the proposed HbA_{1c} estimate in a 10-fold cross-validation test was significantly higher for sparse BG epochs, and was higher or comparable for epochs with high density of measurements. As a result, the proposed approach achieves the main design objective, i.e., improvement of accuracy and robustness of HbA_{1c} prediction.

References

[1] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, and B. Malanda, "Idf diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes research and clinical practice*, vol. 138, pp. 271–281, 2018.

[2] H. F. Bunn, K. H. Gabbay, and P. M. Gallop, "The glycosylation of hemoglobin: relevance to diabetes mellitus," *Science*, vol. 200, no. 4337, pp. 21–27, 1978.

[3] M. Nordwall, M. Abrahamsson, M. Dhir, M. Fredrikson, J. Ludvigsson, and H. J. Arnqvist, "Impact of HbA1c, followed from onset of type 1 diabetes, on the development of severe retinopathy and nephropathy: The viss study (vascular diabetic complications in Southeast Sweden)," *Diabetes Care*, 2015.

[4] R. J. Koenig, C. M. Peterson, R. L. Jones, C. Saudek, M. Lehrman, and A. Cerami, "Correlation of glucose regulation and hemoglobin aic in diabetes mellitus," *New England Journal of Medicine*, vol. 295, no. 8, pp. 417–420, 1976.

[5] P. A. Svendsen, T. Lauritzen, U. Søegaard, and J. Nerup, "Glycosylated haemoglobin and steady-state mean blood glucose concentration in type 1 (insulin-dependent) diabetes," *Diabetologia*, vol. 23, no. 5, pp. 403–405, 1982.

[6] C. L. Rohlfing, H.-M. Wiedmeyer, R. R. Little, J. D. England, A. Tennill, and D. E. Goldstein, "Defining the relationship between plasma glucose and hba1c: analysis of glucose profiles and hba1c in the diabetes control and complications trial," *Diabetes care*, vol. 25, no. 2, pp. 275–278, 2002.

[7] D. M. Nathan, J. Kuenen, R. Borg, H. Zheng, D. Schoenfeld, and R. J. Heine, "Translating the A1C assay into estimated average glucose values," *Diabetes Care*, 2008.

[8] B. P. Kovatchev, F. Flacke, J. Sieber, and M. D. Breton, "Accuracy and Robustness of Dynamical Tracking of Average Glycemia (A1c) to Provide Real-Time Estimation of Hemoglobin A1c Using Routine Self-Monitored Blood Glucose Data," *Diabetes Technology & Therapeutics*, 2014.

[9] B. P. Kovatchev and M. D. Breton, "Hemoglobin A1c and Self-Monitored Average Glucose :Validation of the Dynamical Tracking eA1c Algorithm in Type 1 Diabetes," *Journal of Diabetes Science and Technology*, 2016.

[10] R. J. McCarter, J. M. Hempe, R. Gomez, and S. A. Chalew, "Biological variation in hba1c predicts risk of retinopathy and nephropathy in type 1 diabetes," *Diabetes care*, vol. 27, no. 6, pp. 1259–1264, 2004.

[11] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[13] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[14] R. Hillson, *Diabetes care: a practical manual*. OUP Oxford, 2015.

[15] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[18] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, "Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes," *Journal of medical Internet research*, vol. 21, no. 5, p. e11030, 2019.

[21] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," *arXiv preprint arXiv:1506.00019*, 2015.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[24] S. Garg and L. Jovanovic, "Relationship of fasting and hourly blood glucose levels to hba1c values: safety, accuracy, and improvements in glucose profiles obtained using a 7-day continuous glucose sensor," *Diabetes Care*, vol. 29, no. 12, pp. 2644–2649, 2006.

[25] M. Grady, D. Campbell, K. MacLeod, and A. Srinivasan, "Evaluation of a blood glucose monitoring system with automatic high-and low-pattern recognition software in insulin-using patients: pattern detection and patient-reported insights," *Journal of diabetes science and technology*, vol. 7, no. 4, pp. 970–978, 2013.

[26] B. P. Kovatchev, "Metrics for glycaemic control: from HbA1c to continuous glucose monitoring," *Nature Reviews Endocrinology*, 2017.

[27] J. Li and H. Liu, "Challenges of feature selection for big data analytics," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 9–15, 2017.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[29] D. S. Group *et al.*, "Training in flexible, intensive insulin management to enable dietary freedom in people with type 1 diabetes: dose adjustment for normal eating (dafne) randomised controlled trial," *BMJ: British medical journal*, vol. 325, no. 7367, p. 746, 2002.

[30] O. Heylen, S. Van Neyghem, S. Exterbille, C. Wehlou, F. Gorus, and I. Weets, "Evaluation of the sebia capillarys 2 flex piercing for the measurement of hba1c on venous and capillary blood samples," *American journal of clinical pathology*, vol. 141, no. 6, pp. 867–877, 2014.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] J. Zhou, Y. Mo, H. Li, X. Ran, W. Yang, Q. Li, Y. Peng, Y. Li, X. Gao, X. Luan *et al.*, "Relationship between hba1c and continuous glucose monitoring in chinese population: a multicenter study," *PloS one*, vol. 8, no. 12, p. e83827, 2013.

[33] R. W. Beck, C. G. Connor, D. M. Mullen, D. M. Wesley, and R. M. Bergenstal, "The fallacy of average: how using hba1c alone to assess glycemic control can be misleading," *Diabetes care*, vol. 40, no. 8, pp. 994–999, 2017.