A Multitask Learning Approach to Personalized Blood Glucose Prediction

John Daniels, *Student Member, IEEE*, Pau Herrero, *Member, IEEE*, and Pantelis Georgiou, *Senior Member, IEEE*

***Abstract*—Blood glucose prediction algorithms are key tools in the development of decision support systems and closed-loop insulin delivery systems for blood glucose control in diabetes. Deep learning models have provided leading results among machine learning algorithms to date in glucose prediction. However these models typically require large amounts of data to obtain best personalised glucose prediction results. Multitask learning facilitates an approach for leveraging data from multiple subjects while still learning accurate personalised models. In this work we present results comparing the effectiveness of multitask learning over sequential transfer learning, and learning only on subject-specific data with neural network and support vector regression. The multitask learning approach shows consistent leading performance in predictive metrics at both short-term and long-term prediction horizons. We obtain a predictive accuracy (RMSE) of 18.8±2.3, 25.3±2.9, 31.8±3.9, 41.2±4.5, 47.2±4.6 mg/dL at 30, 45, 60, 90, and 120 min prediction horizons respectively, with at least 93% clinically acceptable predictions using the Clarke Error Grid (EGA) at each prediction horizon. We also identify relevant prior information such as glycaemic variability that can be incorporated to improve predictive performance at long-term prediction horizons. Furthermore, we show consistent performance - $\leq$5% change in both RMSE and EGA (Zone A) - in rare cases of adverse glycaemic events with 1-6 weeks of training data. In conclusion, a multitask approach can allow for deploying personalised models even with significantly less subject-specific data without compromising performance.**

***Index Terms*—Glucose prediction, multitask learning, transfer learning, deep learning, neural network, type 1 diabetes.**

I. Introduction

DIABETES is one of the largest non-communicable diseases in modern society. It is noted that more than 400 million people around the world are living with some variant of diabetes and this number is likely to rise in the foreseeable future [1]. As a result of the scale involved, managing this chronic disease effectively is a pressing health issue that will involve developing solutions to empower individuals with diabetes to manage their condition.

Tight glycaemic control is the primary method of management for diabetes [2]. Digital approaches such as decision support systems and closed-loop insulin delivery (artificial pancreas) systems are currently the focus of enabling tight glycaemic control and alleviating the burden of diabetes management [3], [4].

The proliferation of low power wearable devices and biosensors such as continuous glucose monitors (CGM) and activity bands enable the development of these diabetes management systems for ambulatory monitoring and control [5]. Specifically, the data produced from these sensors and insulin pumps, along with self-reported logs - such as meals and exercise - can be used to predict glucose trajectories [6]. This facilitates the development of methodical strategies for undertaking pre-emptive actions for minimising or avoiding adverse glycaemic events.

Recent works in the literature have demonstrated that deep learning [7] performs better relative to traditional machine learning approaches that usually rely on feature engineering [8]. An advantage of the deep learning approach is that optimal features can be learned to develop personalised models for each individual [9]. However, deep learning models typically require large amounts of data in order to achieve accurate performance [10], [11].

A common challenge in the area of glucose prediction is that large amounts of subject-specific data are expensive and difficult to collect. This hampers the earlier deployment of personalised models which would aid in providing necessary interventions earlier to improve glucose control. Leveraging population data in existing datasets to address this scarcity is further complicated by inter-individual variability [12], [13]. Studies have suggested that differences exist in the complex glucose dynamics of individuals and clustered groups. As a result, there is a possibility that the performance of personalised models can be hampered when consideration is not given to the prior background information such as glycaemic variability [14].

This paper introduces an end-to-end deep multitask approach to developing personalised models while overcoming the issue of inter-individual variability. The main aim is to study effective learning from population data in the development of personalised prediction models. We focus on the following areas in this work.

Manuscript received January 11, 2021; revised June 6, 2021; accepted July 18, 2021. Date of publication July 27, 2021; date of current version January 5, 2022. This work was supported by the Engineering and Physical Sciences Research Council, funded under ARISES project EP/P00993X/1. *(Corresponding author: John Daniels.)*

The authors are with the Department of Electrical, and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: john.daniels11@imperial.ac.uk; p.herrero-vinias@imperial.ac.uk; pantelis@imperial.ac.uk).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2021.3100558, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2021.3100558

- We investigate the effect of transfer learning approaches for blood glucose prediction across different prediction horizons.
- We investigate the effect of incorporating background information such as glycaemic variability on glucose prediction performance.
- We investigate the impact of training data size on the performance of multitask models.

This work builds on preliminary work that demonstrate the performance of multitask learning for personalised blood glucose prediction at 30-min and 60-min prediction horizons [15]. We now compare this approach to inductive transfer against the popular sequential transfer learning (TL) approach that involves fine-tuning, and single-task learning (STL) models trained solely on subject-specific data. We also include support vector regression (SVR) as an additional traditional baseline approach for comparison.

The rest of the paper is organised as follows: Section II discusses the related work in the literature regarding transfer learning and leveraging population data for glucose prediction. Section III details the methods involved in this work; this covers the architecture and training of the deep multitask networks. Section IV details the methods - dataset and models - used in carrying out the experiments. Section V provides a presentation of the results in the experiments. Section VI is a discussion of the results in context of current knowledge, and limitations to be addressed in future work.

II. Related Work

Transfer learning is a field that considers leveraging knowledge from previous experience to improve learning for a related particular task. Formally, a task $\mathcal{T}$ comprises a label space $\mathcal{Y}$ and a predictive function f that is learnt from available data. The available data is associated with a domain $\mathcal{D}$, a space that comprises the input feature space X and the output Y [16]. Each task and domain associated with previous experience is termed a source domain and source task $\{(\mathcal{D}_S, \mathcal{T}_S)\}$ and the particular task and domain of interest is termed target domain and target task,$\{(\mathcal{D}_T, \mathcal{T}_T)\}$. Therefore in transfer learning, $\{(\mathcal{D}_S, \mathcal{T}_S)\}$ is leveraged along with $\{(\mathcal{D}_T, \mathcal{T}_T)\}$ to improve the ability to learn the target predictive function f_T.

The success of transfer learning has been noted in the fields of computer vision, digital imaging, natural language processing, as well as other areas of healthcare [16]–[20]. This typically considers using a well sourced large dataset (eg. ImageNet) in order to pre-train models, before subsequently fine-tuning models on data from the target task [10]. In order to be successful, the tasks are usually assumed to be related in some sense.

In the field of blood glucose prediction, transfer learning is not a well studied approach despite the success of data-driven methods in glucose prediction [8]. This could be attributed to the lack of large publicly available datasets in the field. For most studies that have employed deep learning, results have provided neither a consensus nor a detailed analysis on the methods and benefits of information transfer from source tasks to target tasks.

For short-term predictions (PH $\leq$ 60 minutes), the results on transfer learning are mixed. *Bhimireddy et al.* [21] employ a sequence-to-sequence network and do not report improved performance with transfer learning. On the other hand, *Rubin-Falcone et al.* [22] pre-train their models in a two-stage process; the authors first train on a large private dataset of 100 subjects before subsequently training the general model on the six subjects in the first cohort of OhioT1DM subjects and then fine-tuning on the latter cohort. This yields a 4% improvement in average predictive performance in terms of root mean square error (RMSE).

Further works have also looked at variant transfer learning approaches to leverage population data [23]–[25]. *Hameed et al.* [23] use knowledge distillation [26] in order to learn models that leverage large datasets. In this case, a teacher model is used to learn an initial model and the student model is subsequently trained in the target domain and the outputs - soft prediction labels - from the teacher model. However, this approach was unable to improve performance over the student only approach trained solely on subject-specific data. Other approaches [24], [25] aim to learn from different metabolic classifications i.e. Type 2 diabetes and Type 1 diabetes. The approach of *Gu et al.* [24] is positive, although the model was only tested on one day of glucose data and the authors are unclear on whether baseline models were trained on the same data to evaluate the benefit of the transfer approach. A domain adversarial learning approach enables learning features common to each group, thereby improving generalisation and performance [25].

Kushner et. al [27] study the benefit of sequential transfer learning at long-term prediction horizons beyond 60 minutes. This work shows some benefit in the average performance of the neural network model on the subjects. However, further analysis suggests subjects with high glycaemic variability did not benefit from transfer learning.

III. Multitask Learning

Multitask learning (MTL) is a type of transfer learning that involves learning multiple tasks simultaneously in order to improve generalisation [28]. The transfer of information signals occurs in parallel since the models for each task are jointly learned. This differs from the typical transfer learning approach as shown in Fig. 1, which can best be described as sequential transfer, where the models are pre-trained on the data from the source domain towards source tasks before training on data from the target domain towards the target task. This approach can help to overcome the issue of inter-individual variability that exists in developing optimal personalised models within a feasible timeline.

The architecture and training protocol for multitask learning is different to the single-task (STL) and fine-tuning (TL) approach. In this setting, the models are trained from random initialisation similar to STL, however the personalised models are trained jointly in order to facilitate transfer.

A. Network Architecture

Multitask learning for neural networks is realised by sharing the parameters of the layers between the tasks in the multitask model. The architecture of the multi-task neural networks are detailed in Fig. 2.

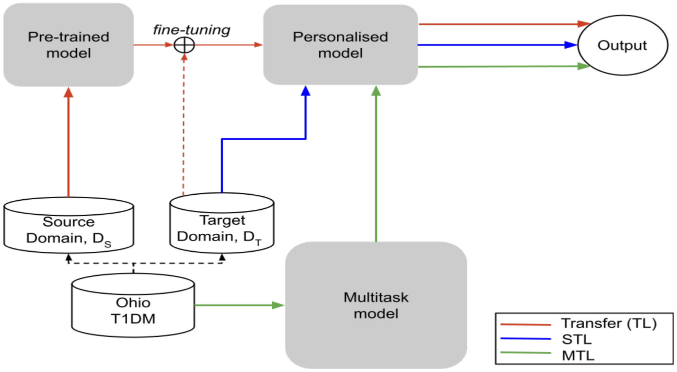


Fig. 1. A simplified illustration of the training steps for each of the learning approaches. This covers subject-specific single-task learning (STL), transfer learning (TL) with fine-tuning, and multitask learning (MTL).

The initial layers of the neural networks are shared between tasks (i.e. subjects), and the final layer of the multitask models are task-specific. As a result, the parameter sharing is achieved by connecting each task-specific to the single branch of initial layers as shown in Fig. 2(a). Sharing more parameters in the model serves as an additional form of regularisation, such that the model is constrained to learn features relevant to all tasks. This benefit underlying multitask learning is conditioned on the tasks being similar so that such features exist. In the scenario where tasks are not similar, model performance can be hampered relative to models trained in a single-task learning approach - this is termed negative transfer.

To overcome the challenge of inter-individual variability that results in tasks being less similar, hence potential negative transfer, the degree of parameter sharing is relaxed. This is shown in Fig. 2(b) where intermediate layers are clustered based on prior information. For this work, we cluster individuals based on their glycaemic variability.

B. Network Training

The training protocol in the multitask learning (MTL) setting differs from the single-task learning (STL) setting. In the single-task setting, given that data during a training session is subject-specific, the training samples can be selected in mini-batches sequentially as typically expected in supervised learning for time-series regression.

However, in the multitask setting, the training samples contain samples for the number of tasks (subjects) present. Consequently, samples in a mini-batch are drawn from a particular individual to train the shared layers and the layers specific to the individual similar to [19]. The loss function, $\mathcal{L}_{MAE}(y, \hat{y})$, used to minimise the error during training is defined below:

$$\mathcal{L}_{MAE}(y, \hat{y}) = \frac{1}{N_{batch}} \sum_{k=1}^{N} |y_k - \hat{y}_k|, \qquad (1)$$

where $\hat{y}$ denotes the predicted results given the historical data and y denotes the reference change in glucose concentration over the relevant glucose prediction, and N_{batch} refers to the number of samples in the mini-batch.

A sample weighting is used as a gating approach in order to ensure that the input corresponding to a subject is used to train layers in the network that pertain to the associated subject. During a forward pass the mini-batch is fed into the network to obtain the predicted glucose values, $\hat{y}$, and determine the loss according to 1. At each iteration, the backpropagated error is used to learn personalised weights of each subject in the task specific layers and eventually learn appropriate weights in shared layers that generalise to all subjects at the same time.

TABLE I
BACKGROUND INFORMATION AND DATA SIZES FOR SUBJECTS IN THE OHIOT1DM DATASET

ID	Gender	Age	Glycaemic Variability (CV)	Training Set Size	Testing Set Size
540	M	20 - 40	Labile (40%)	11947	2884
544	M	40 - 60	Stable (36%)	10623	2704
552	M	20 - 40	Labile (37%)	9080	2352
559	F	40 - 60	Labile (42%)	10796	2514
563	M	40 - 60	Stable (34%)	12124	2570
567	F	20 - 40	Labile (40%)	10877	2377
570	M	40 - 60	Stable (33%)	10982	2745
575	F	40 - 60	Labile (43%)	11866	2590
584	M	40 - 60	Stable (34%)	12150	2653
588	F	40 - 60	Stable (31%)	12640	2791
591	F	40 - 60	Labile (37%)	10847	2760
596	M	60 +	Stable (33%)	10877	2731

IV. METHODS

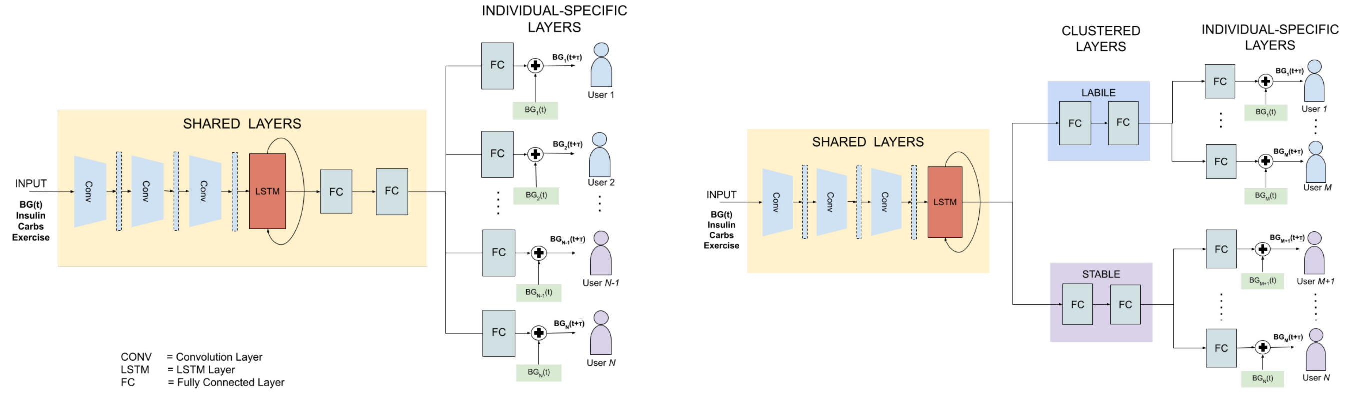
A. Dataset

The dataset used in this study is referred to as the OhioT1DM dataset [29], and will be referred to as such from here on. This dataset was obtained under the Data use Agreement (DUA) between Ohio University and Pennsylvania State University. The OhioT1DM dataset is updated on 2020 and comprises 12 subjects with Type 1 diabetes (T1D) monitored in free-living conditions over a period of 8 weeks.

As seen in Table I above, the dataset contains 7 male subjects and 5 female subjects. In terms of age, all subjects are adults; with subjects grouped as young adults (20-40 years), middle aged adults (40-60 years), and old adults (60+ years). The glycaemic variability [14] is determined with the training data samples and is formulated through the coefficient of variation (CV) as denoted below:

$$CV = \frac{\sigma}{\mu} \times 100\%, \qquad (2)$$

where CV is the coefficient of variation, σ is the standard deviation of the glucose concentration levels, and μ is the mean of the glucose concentration levels. A subject is classified as labile if CV $>$ 36%, and stable otherwise. This threshold represents an increased incidence of hypoglycaemia in the glucose profile and is further supported in the literature [30]. The percentage time spent in hypoglycaemia for all individuals is provided in Supplementary Table I and a comparison of stable and labile groups is shown in Supplementary Fig. I.



(a) MTL

(b) MTL-GV

Fig. 2. (a) The multitask network architecture shows the convolutional and recurrent layers are shared between all subjects. The fully connected are shared to varying degrees if clusters are specified. (b) The multitask network architecture shows the convolutional and recurrent layers are shared between all subjects. The fully connected are shared to varying degrees if clusters are specified.

The subjects are provided with a Medtronic Enlite Continuous Glucose Monitoring (CGM) devices along with one of a Basis Peak Band (Intel Corp. Santa Clara, CA, US) or Empatica Embrace (Empatica, Inc., Boston, MA, US). The CGM measures interstitial glucose concentration levels at 5 minute intervals. The Basis Peak measures values of the skin temperature, skin conductance, heart rate, and step count (this has been aggregated over 5-minute intervals). On the other hand, the Empatica Embrace measures skin conductance, skin temperature, and acceleration magnitude (aggregated over 5-minute intervals). In addition to the physiological signals, subjects provide self-reported assessments such as meal intake, insulin, exercise, sleep, stressors, work, and sleep.

B. Preprocessing

Prior to training and testing, the data undergoes processing to facilitate effective learning. We first prepare the real-world data with normalisation and imputation.

In order to evaluate the models we mainly consider four features that are both prominent for the task and consistent for each subject: glucose concentration levels, insulin, meals, and exercise. The first approach is to synchronize the data entries for each modality, using the CGM timestamp as the reference timestamp. A sliding window is used to extract 2-hour sequences of historical data for each prediction, selected based on analysis in previous work [31].

The reported exercise is first transformed from the initial range of 1-10 to a binary representation denoting the presence (1) or absence (0) of exercise. We normalise the remaining variables in order to facilitate easier learning.

The non-trivial issue of missingness is next addressed in processing this dataset [32]. Missingness is present in both the physiological variables (CGM) and self-reported data and these are handled differently. In order to overcome this we implement data imputation methods in the training set and the testing set.

The glucose concentration values that are missing in this dataset are assumed to be missing at random. This means that the lack of data at these timestamps can be attributed to random circumstances such as changing sensor, power aberration, and communication failure among others.

In the training set, the gaps in the glucose concentration levels are imputed using a linear interpolation. This assumes that the missing data is adequately explained with a linear relationship between adjacent glucose levels. For samples where the input sequence of the glucose concentration values contains more than one hour (12 samples) of imputed values, the sequence is discarded. This is done to avoid learning artifacts and incorrect trends since it is difficult to determine the long-term effect of missing information such as meals on missing glucose values. Regarding the self-reported data, that is regarded as missing not at random. Consequently, the assumption made is that a report not made at a particular timestamp, is due to an absence of the activity. As such the gaps in data for insulin, meals, and exercise are imputed with zero.

For accurate evaluation of the performance of the model on all test points in the test set, using only interpolation at test time is not appropriate since we may be using unknown future values. We employ different modes of extrapolation in order to impute missing CGM values in the input sequence as detailed in Fig. 3.

C. Glucose Prediction Models

We describe the glucose prediction models used in the experiments. These are support vector regression and a deep learning model (convolutional recurrent neural network).

1) Support Vector Regression (SVR): The support vector regression has been shown in the literature [33], [34] to provide competitive performance in the area of glucose prediction. This used a radial basis function (RBF) as the kernel and provides a good benchmark given the ability to perform well with small datasets. The SVR model is trained only on subject-specific data. SVR is developed using Scikit-learn v0.21.3 Python library [35].

2) Deep Learning: Many deep learning architectures exist in the literature [9]. Although it is difficult to establish a standout state-of-the-art approach, most approaches are based on

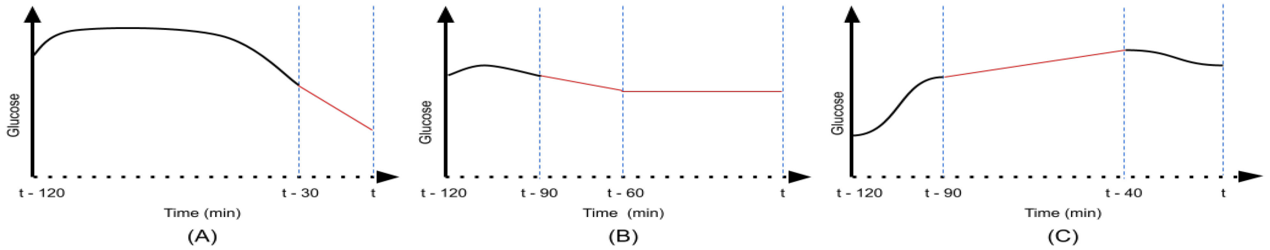


Fig. 3. A visualisation of the imputation methods employed in this work. In (a) the 2-hour input sequence has up to 30 minutes of recent values missing and is imputed with linear extrapolation. (b) shows the imputation scheme during testing for longer than 30 minutes of recent values missing (zero-order hold). Finally (c) shows the imputation scheme when the missing values of the input sequence are located between real values (linear interpolation).

recurrent networks. We use the convolutional recurrent neural network (CRNN) architecture, a 6-layer end-to-end learning framework developed by *Li et. al* [31] for investigating the different learning approaches detailed below. The models are developed with Python 3.6 and Keras v2.2.2 [36] and trained using a NVIDIA GTX 1050.

The initial stage of the CRNN model involves 3 causal convolutional layers - designed with 8, 16, and 32 filters - that are used for extracting features from the inputs.

These features are then fed into a recurrent LSTM layer which is suitable for modelling sequential data. The LSTM is a variant of recurrent neural network (RNN) that handles the issue of vanishing gradients that hampers learning in vanilla RNNs [37]. This is done with additional gates (i.e. the input, output, and forget gates) in the LSTM cell to add or remove information during learning, thus regulating the cell state [37]. The output of this stage is then processed by 2 fully connected layers to obtain the output of the model, $\hat{y}_t$.

$$BG_{(t+\tau)} = BG_t + \hat{y}_t \tag{3}$$

Finally, the model output is added to the present glucose value, BG_t, to obtain the final predicted glucose value, $BG_{t+\tau}$.

2.1) **Single-Task Learning (STL)**: In the single task learning setting, the CRNN model is trained from random initialisation solely on data from the distinct subject.

2.2) **Transfer Learning (TL)**: In the transfer learning setting, the model is first pre-trained on data from the other subjects. The weights in all layers are frozen except the final layer. The model is then fine-tuned on data from the target subject.

2.3) **Multitask Learning (MTL)**: In the multitask setting, the weights in the model are trained from random initialisation, similar to the STL learning approach. Models are trained jointly using all 12 subjects.

2.4) **Multitask Learning (Glycaemic Variability) (MTL-GV)**: In this multitask setting, the training approach the same as the MTL training approach and models are trained jointly using all 12 subjects. The difference lies in the network architecture as seen in Fig. 2.

During training, we split the last 10% of the training data as the validation set. We set the number of epochs to 200 and implement early stopping with a patience of 20 epochs ($\Delta_{min} = 1 \times 10^{-4}$) to terminate training when validation loss is no longer improving.

The optimised hyperparameters for the various glucose models are presented in Table IV in the Appendix below.

D. Criteria for Assessment

We employ multiple criteria to comprehensively evaluate model performance in the following areas:

- Predictive accuracy in terms of the magnitude of error from the reference values in the dataset.
- Temporal gain in terms of prediction horizon relative to reference values in the dataset.
- Clinical significance of errors to understand subsequent use in diabetes management systems, particularly in extreme adverse glycaemic event regions.

The predictive performance of the model is primarily evaluated by the root-mean-square error (RMSE) and mean absolute error (MAE). The effective prediction horizon (PH_{eff}) is used to evaluate the temporal gain in forecasting a glucose concentration value. This is determined using cross-correlation between the predicted and reference glucose concentration levels.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y(k) - \hat{y}(k))^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y(k) - \hat{y}(k)| \tag{5}$$

$$\begin{aligned} PH_{eff} &= PH - \tau_{delay} \\ &= PH - \arg\max_{\tau}(y(k) \star \hat{y}(k)) \end{aligned} \tag{6}$$

where $\hat{y}(k)$ denotes the predicted glucose level for a given sample, k. $y(k)$ denotes the reference glucose measurement, N refers to the number of samples in the testing set. PH refers to the expected prediction horizon, and τ_{delay} refers to the lag in predicted glucose level.

The Clarke Error Grid Analysis (EGA) was originally developed to quantify the clinical accuracy of current blood glucose estimates against reference blood glucose values [38]. We adopt this approach, as done in literature, to evaluate the clinical accuracy of glucose forecasting algorithms as shown in Fig. 4 [22], [27]. The graph is demarcated into five zones, labelled A-E, that represent increasing severity of errors due to misestimation of predicted glucose concentration levels as follows:

TABLE II
Comparison of the Performance Metrics of Deep Multitask Learning Models Against Conventionally Trained Deep Neural Networks and SVR Models at Different Prediction Horizons (Best Result Highlighted in BOLD)

Prediction Horizon (min)	Metric	CRNN				SVR
		MTL	MTL-GV	TL	STL	
30	RMSE (mg/dL)	**18.8 ± 2.3***	**18.8 ± 2.8**	19.2 ± 2.2	20.6 ± 2.6*	19.2 ± 2.7
	MAE (mg/dL)	**13.2 ± 1.6**	**13.2 ± 1.5**	13.4 ± 1.5	14.8 ± 2.1*	13.5 ± 1.7
	PH_{eff} (min)	13.8 ± 5.1	13.3 ± 5.9	13.3 ± 4.3	10.6 ± 5.2	**14.6 ± 5.6***
	EGA (%)	99.1 ± 0.7	**99.2 ± 0.5**	99.1 ± 0.6	98.6 ± 0.7*	98.9 ± 0.8
45	RMSE (mg/dL)	**25.3 ± 2.9***	25.9 ± 3.1	26.5 ± 3.0	26.8 ± 3.5	26.5 ± 4.3
	MAE (mg/dL)	**18.2 ± 2.2**†	19.0 ± 2.4	18.9 ± 2.2	19.5 ± 2.7	19.7 ± 3.5
	PH_{eff} (min)	**19.6 ± 6.3**	17.5 ± 8.5	15.8 ± 4.0	14.2 ± 7.6	19.2 ± 2.5†
	EGA (%)	97.9 ± 1.5	97.5 ± 2.0	**98.1 ± 1.3**	97.8 ± 1.7	97.7 ± 1.7
60	RMSE (mg/dL)	**31.8 ± 3.9***	32.3 ± 3.9*	33.0 ± 3.7	33.9 ± 4.3	32.6 ± 4.0
	MAE (mg/dL)	**23.4 ± 3.0***	23.6 ± 3.0†	24.4 ± 2.8	25.2 ± 3.5	24.0 ± 3.2
	PH_{eff} (min)	**20.4 ± 8.3**	17.5 ± 8.8	14.2 ± 5.3	12.9 ± 8.3	12.9 ± 6.9*
	EGA (%)	96.8 ± 2.1	**97.1 ± 2.0***	96.6 ± 2.3	96.2 ± 2.8	96.6 ± 2.2
90	RMSE (mg/dL)	**41.2 ± 4.5***	41.5 ± 4.3*	43.2 ± 4.5	43.1 ± 5.4	42.6 ± 4.8
	MAE (mg/dL)	**31.1 ± 3.7***	31.2 ± 3.4†	32.6 ± 3.4	32.7 ± 4.1	32.5 ± 4.1
	PH_{eff} (min)	21.2 ± 9.8*	20.0 ± 11.9	15.0 ± 7.6	18.7 ± 12.3	**32.1 ± 1.8***
	EGA (%)	**95.0 ± 3.0**	**95.0 ± 2.9***	94.6 ± 2.8	94.5 ± 3.0	94.9 ± 3.0
120	RMSE (mg/dL)	48.0 ± 5.2*	**47.2 ± 4.6***	49.3 ± 4.9	49.0 ± 5.4	48.0 ± 5.1
	MAE (mg/dL)	37.1 ± 4.1*	**36.5 ± 3.8***	38.3 ± 3.7	37.9 ± 4.1	37.5 ± 4.0
	PH_{eff} (min)	26.8 ± 13.8*	26.3 ± 13.1*	15.8 ± 7.3	14.4 ± 1.9	**27.3 ± 14.2***
	EGA (%)	93.7 ± 3.0*	**93.8 ± 2.8***	92.8 ± 2.9	93.1 ± 3.3	93.1 ± 3.6

Statistically significant compared to TL with p-value $<.013$ (*Paired t-test; †Wilcoxon).

TABLE III
Total Number of Samples at Different Training Set Sizes and Associated Samples in the Hypoglycaemia Region

Duration	Number of Training Samples	
	All	Hypoglycaemia
6 weeks	134790 (9080 - 12640)	4962 (110 - 1034)
4 weeks	88820 (5690 - 8380)	3120 (55 - 668)
2 weeks	44712 (2945 - 3986)	1367 (50 - 251)
1 week	22470 (1454 - 2148)	732 (8 - 194)
3.5 days	10710 (574 - 1093)	455 (5 - 91)

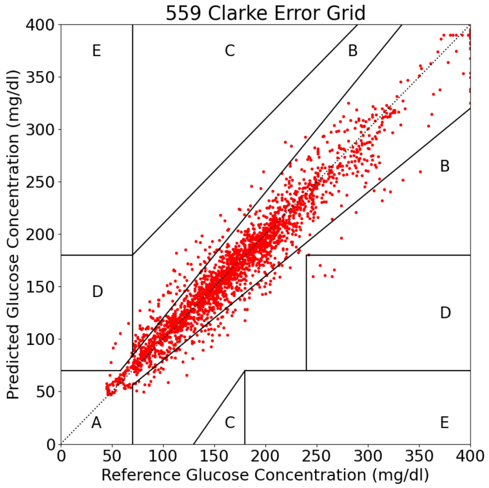


Fig. 4. Clarke Error Grid showing the location of points in the various zones of safety for Subject 559. This illustrates the clinical relevance of errors by the MTL model at a 30-minute prediction horizon.

- Zone A: Predicted values in this region lie within 20% of the reference CGM values when CGM ≥ 70 mg/dL, and predicted CGM values are no more than 70 mg/dL during hypoglycaemia.
- Zone B: Errors of predicted values in this region fall outside the 20% error, however, any resulting standard treatment could be incorrect but uncritical.
- Zone C: Predicted values in this region could result in unnecessary treatment.
- Zone D: Predicted values in this region point to a potentially harmful adverse glycaemic event (hyperglycaemia or hypoglycaemia) that has gone undetected.
- Zone E: Predicted values in this region if acted on could lead to the opposite corrective action being undertaken to treat an adverse glycaemic event.

E. Statistical Analysis

For determining the statistical significance of differences between model performances, we first perform preliminary test for normality using the Shapiro-Wilk test. We use a paired t-test if normality is accepted, and a Wilcoxon signed-rank test when normality is rejected. Significance level is set at p-value $<.05$. For multiple pairwise comparisons, we adjust the significance level to p-value $<.013$ using Bonferroni correction.

V. Results

In this section we detail the performance of the CRNN model trained with the multitask learning approach against the CRNN model trained with a single-task learning and transfer learning approach as well as the SVR baseline method. The various algorithms and training approaches are evaluated on the OhioT1DM dataset.

TABLE IV
HYPERPARAMETER SEARCH SPACE AND CONFIGURATION

Hyperparameter	Value Range	Prediction Horizon				
		30	45	60	90	120
	CRNN					
Kernel size	CONV {1, 2, 4, 8}	4	4	2	1	1
Number of units	LSTM {8, 16, 32, 64}	32	32	32	64	32
	FC (1) {64, 128, 256, 512}	256	128	256	64	512
	FC (2) {8, 16, 32, 64}	16	16	32	64	16
Dropout rate	{0.10,..., 0.90}					
	CONV	0.40	0.36	0.50	0.23	0.52
	LSTM	0.17	0.58	0.74	0.54	0.18
	FC	0.42	0.33	0.41	0.36	0.24
Learning rate	$\{1 \times 10^{-5}$,..., $1 \times 10^{-2}\}$					
	STL	3.7×10^{-4}	1.2×10^{-3}	3.6×10^{-4}	3.0×10^{-3}	7.0×10^{-4}
	TL	4.3×10^{-5}	4.3×10^{-5}	2.9×10^{-5}	2.5×10^{-3}	1.3×10^{-4}
	MTL	6.0×10^{-4}	5.0×10^{-4}	1.1×10^{-3}	5.4×10^{-5}	1.0×10^{-3}
	MTL-GV	4.8×10^{-4}	2.0×10^{-3}	6.7×10^{-4}	7.5×10^{-4}	3.8×10^{-4}
Batch size	{64, 128, 256, 512}					
	STL	256	128	256	256	256
	TL	256	128	64	256	512
	MTL	64	64	128	64	256
	MTL-GV	128	512	64	64	64
	SVR					
C	{0.1,...,1000}	237	610	980	960	170
Gamma	$\{1 \times 10^{-4}$,..., $1 \times 10^{-2}\}$	1.4×10^{-3}	9.7×10^{-4}	4.0×10^{-4}	1.6×10^{-3}	1.1×10^{-4}
Epsilon	$\{1 \times 10^{-4}$,..., 1}	1.4×10^{-2}	2.0×10^{-4}	3.1×10^{-4}	1.9×10^{-1}	2.6×10^{-1}

A. Performance Comparison Across Prediction Horizons

In this experiment, we investigate the performance of the multitask learning approaches against current learning approaches and methods at prediction horizons. We evaluate the performance of MTL and MTL-GV against TL, STL, and SVR approaches for 30-120 min. MTL-GV incorporates prior information on subject glycaemic variability whereas MTL does not incorporate any prior information in the network architecture. The results for this are shown in Table II.

At the short-term prediction horizon (<60 min), multitask learning shows the best performance in terms of both predictive accuracy metrics. MTL and MTL-GV showed the best mean RMSE (18.8 mg/dL) and MAE (13.2 mg/dL) compared to TL (+0.4 mg/dL and +0.2 mg/dL), STL (+1.8 mg/dL and +1.6 mg/dL), and SVR (+0.4 mg/dL and +0.3 mg/dL) models. Compared to the conventional approach of transfer learning by finetuning (TL), MTL reveals a significant improvement (*p*-value $<$. 013) in terms of RMSE at 30 min and in terms of RMSE and MAE at 45 min. However, the improvement in metrics compared to MTL-GV are non-significant.

The trend of MTL demonstrating the best performance remains as the prediction horizon increases. At long term predictions (≥60 min), the MTL and MTL-GV models generally outperform the TL, STL and SVR models. Both MTL and MTL-GV models reveal significant improvement compared to TL models at long-term predictions in terms of both RMSE and MAE. At 120 minutes, MTL performs at least as well as the SVR models in terms of RMSE and is slightly better (-0.4 mg/dL) in terms of MAE. Further results on model predictive accuracy at specific regions (hypoglycaemia, euglycaemia, and hyperglycaemia) are reported in Supplementary Table III.

Multitask learning, in regard to clinical accuracy, shows a comparable performance with other models. Compared to TL models, MTL only shows a significant improvement in performance at 120 min. On other hand, MTL-GV models show a significant improvement from 60 min onward. For the prediction horizons studied, the MTL and MTL-GV approaches maintain at least 93% of predictions within Zone A or Zone B.

In terms of temporal gain, the MTL models show a higher temporal gain relative to other conventionally trained models (TL and STL). However, the results of the temporal gain of multitask learning models (MTL and MTL-GV) are mixed compared to SVR models. MTL and MTL-GV models show a higher temporal gain at 45 min and 60 min, but lower temporal gain at 30 min, 90 min and 120 min.

B. Incorporating Prior Information on Glycaemic Variability

We also examine the effect of incorporating prior information. In this study we incorporate information on glycaemic variability in the architecture of the multitask neural network (MTL-GV).

The effect of clustering layers does not lead to an improvement in predictive accuracy until 120 min. As seen in Table II, MTL consistently demonstrates the best predictive accuracy at these prediction horizons, with significant improvements over MTL-GV in terms of RMSE and MAE at 45 min and 60 min (p-value $<.05$). On the other hand, the predictive accuracy of multitask models are improved at 120 min when incorporating glycaemic variability (MTL-GV), in terms of RMSE and MAE (p-value $<.05$), over no specified clustering (MTL). The temporal gain is observed to be affected as this is lower for MTL-GV models compared to MTL models across all prediction horizons.

The Clarke Error Grid Analysis (EGA), which focuses on the percentage of samples in the safe zones (Zone A and Zone B), shows that clustering provides a slight increase in EGA (Zone A+B) at 30 min (+0.1%), 60 min (+0.3%) and 120 min (+0.1%), a decrease at 45 min (-0.4%) and no change at 90 min. MTL-GV generally improves EGA (Zone A+B) of subjects with high glycaemic variability over MTL at 120 mins. The complete EGA metrics for all zones are reported in Supplementary Table II.

C. Impact of Training Data Size

As noted in earlier sections, most public datasets available and suitable for glucose prediction are typically small in size. This has to be considered when deploying these models in diabetes management systems. This experiment investigates the benefit of a multitask learning approach for different training set sample sizes. The prediction horizon is set at 30 minutes as it is the typical prediction horizon used in commercial CGM and predictive low glucose suspend (PLGS) systems [39].

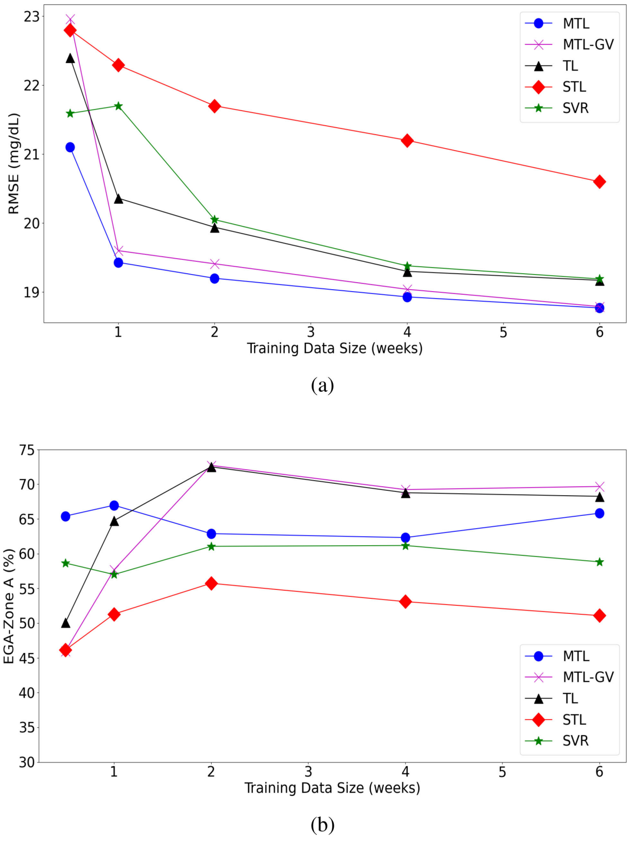
The initial training set size covers 6 weeks. We also evaluate the performance at the following duration periods up to the end of the training set: 4 weeks, 2 weeks, 1 week and 0.5 weeks (3.5 days). Table III shows the total number of training samples for each duration and the range of training samples for individuals. This table also shows the number of training examples in the hypoglycaemia region for the associated duration.

An important aspect is the performance of these models in terms of clinical accuracy as the number training set size reduces. As seen in the first experiment, the clinical accuracy for all models in the safe regions is generally high i.e. Zone A-B $\geq$98%. As seen in Table III, the number of training points in the hypoglycaemia region (CGM $\leq$ 70 mg/dL) are relatively scarce (3-4%), but very consequential for applications such as PLGS systems. The guidance for hypoglycaemia treatment is to ingest *rescue carbohydrates* and/or suspend the insulin basal rate to facilitate recovery of blood glucose concentration levels.

As seen in Fig. 5(a), multitask learning (MTL) provides the best performance in terms of RMSE at all training sizes. This is significant compared to the conventional TL approach (p-value $<.013$). Furthermore, MTL are able to maintain this consistent performance in predictive accuracy when trained on at least 1 week of data (Δ RMSE $\leq$ 5%) from T1DM subjects.

For clinical accuracy, we focus on consistency of model performance in the hypoglycaemia region to determine if early deployment is possible without compromising performance. At all training data sizes, no models reported predictions in Zone E.

Fig. 5(b) shows the clinical relevance of errors in hypoglycaemia at each training set size and highlights the consistency in performance of the MTL approach. MTL approach shows better consistency (Δ_{max} Zone A = -4%) in performance over the various training sizes relative to the MTL-GV approach (Δ_{max} Zone A = -26.8%). In this scenario, MTL-GV models show the best performance when trained with at least 2 weeks of training data from each participant followed by the TL models. However, this performance drops sharply when training data size is further limited - Δ_{max} Zone A = -26.8% for MTL-GV and Δ_{max} Zone A = -22.4% for TL. The best performance in clinical accuracy for hypoglycaemia prediction is obtained by the MTL-GV model when the training data is reduced to the last 2 weeks of samples at 72.7%.



(a)

(b)

Fig. 5. (a) A comparison of the predictive accuracy, in terms of RMSE, for data-driven machine learning models and learning approaches at different training set sizes. The clinical relevance of errors in model prediction for hypoglycaemia are shown in (b).

VI. Discussion and Future Work

Multitask learning facilitates transfer of useful information between subjects. The results detailed in Section V demonstrate that the performance of the personalised models improves with the introduction of population data, however, this benefit is more consistent with a multitask learning approach.

Small datasets can typically result in underperformance of deep learning methods which prompts the use of transfer learning [10]. One potential reason for the improvement in performance with the multitask learning approach, is that parameter sharing may serve as an additional form of regularisation which works to improve the generalisation of model performance on unseen data. This could also explain the high and consistent performance experienced, despite limiting the size of training data, relative to other approaches evaluated.

Multitask learning also allows us to address the possible effect of negative transfer that would lead to a decrease in performance. Clustering layers before final individual-specific layers can reduce the effect of negative transfer. The clustering, however, seems to establish a trade-off between limiting the amount of negative transfer and model performance from reduced parameter sharing. This observation is made given the performance of MTL at different prediction horizons when glycaemic variability is considered. At 120 minutes, the relatedness between all tasks may be reduced which would make negative transfer more prominent over the benefit from parameter sharing. This may also explain the lack of improved performance prediction for PH ≤ 90 min where clustering reduces the degree of parameter sharing (i.e. regularisation) and as a result, model performance is affected.

Assessing the credibility of such models to be deployed in the healthcare domain is important and gaining attention. This credibility can be ascertained through model risk assessment, verification, and validation [40]. The validation and verification of the model can be considered through empirical metrics such as RMSE and MAE. On the other hand, model risk assessment can be evaluated on the EGA and temporal gain. For example, if consistent performance is sought by the clinician, the multitask learning (MTL) approach shows the most consistency in model performance even with reduced training data size. These considerations can give confidence to the clinician to recommend these models in a PLGS system even with limited subject data available.

Limitations in this work exist that can be tackled in future work. One such limitation is that the small number of T1DM subjects in the dataset means we are unable to fully characterise the effect of prior information on multitask performance. As larger open datasets are being made available in the future, we could investigate the impact of incorporating combinations of prior information, such as age and glycaemic variability, on model performance. We could also investigate the effect of other sources of information such as heart rate monitors on multitask learning performance.

VII. Conclusion

Deep learning approaches are increasingly becoming relevant in developing the next-generation of diabetes management tools to aid in diabetes management. Glucose prediction represents a core part of that path, and as a result, effective methodologies are necessary to realise this with data-driven models. Deployment of such personalised models are hampered by the limited size of individual data available. Multitask learning provides an effective approach for leveraging population data to develop personalised glucose prediction models and overcome the challenge of scarce data for training models. Furthermore, incorporating prior information such as glycaemic variability can be beneficial for long-term prediction tools in a multitask setting. Finally, this approach is agnostic to the neural network architecture and can be compatible with other architectures developed in the future. The results from this work suggest multitask learning can facilitate a path for potentially deploying personalised models towards improving glycaemic control on limited individual data.

Appendix Hyperparamter Optimisation

To select hyperparameters for the glucose prediction models, we perform a Bayesian hyperparameter optimization algorithm with Tree of Parzen estimators using CometML [41]. The search space and optimal hyperparameters are shown in Table IV.

Acknowledgment

The authors would like to thank Dr. Cindy Marling and Dr. Razvan Bunescu from Ohio State University for making the dataset available through a Data Use Agreement (DUA).

References

[1] G. Roglic, World Health Organization, “Global Reports on Diabetes,” Geneva, Switzerland: World Health Organization, 2016. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] D. M. Nathan and F. T. D. R. Group, “The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: Overview,” *Diabetes Care*, vol. 37, no. 1, pp. 9–16, Jan. 2014.

[3] R. Ramli, M. Reddy, and N. Oliver, “Artificial pancreas: Current progress and future outlook in the treatment of type 1 diabetes,” *Drugs*, vol. 79, no. 10, pp. 1089–1101, Jul. 2019.

[4] N. S. Tyler and P. G. Jacobs, “Artificial intelligence in decision support systems for type 1 diabetes,” *Sensors*, vol. 20, no. 11, Jan. 2020, Art. no. 3214, Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/20/11/3214

[5] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, “Wearable continuous glucose monitoring sensors: A revolution in diabetes treatment,” *Electronics*, vol. 6, no. 3, pp. 65–81, Sep. 2017. [Online]. Available: http://www.mdpi.com/2079-9292/6/3/65

[6] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, “A review of personalized blood glucose prediction strategies for T1DM patients,” *Int. J. Numer. Methods Biomed. Eng.*, vol. 33, no. 6, Jun. 2017, Art. no. e2833.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[8] A. Z. Woldaregay *et al.*, “Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes,” *Artif. Intell. Med.*, vol. 98, pp. 109–134, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365717306218

[9] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Deep learning for diabetes: A systematic review,” *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2744–2757, Jul. 2021.

[10] M. Huh, P. Agrawal, and A. A. Efros, “What makes ImageNet good for transfer learning?,” Dec. 2016, *arxiv:1608.08614*.

[11] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 843–852.

[12] X. Yu *et al.*, “Online glucose prediction using computationally efficient sparse Kernel filtering algorithms in type-1 diabetes,” *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 1, pp. 3–15, Jan. 2020.

[13] I. Hajizadeh *et al.*, “Incorporating unannounced meals and exercise in adaptive learning of personalized models for multivariable artificial pancreas systems,” *J. Diabetes Sci. Technol.*, vol. 12, no. 5, pp. 953–966, Sep. 2018.

[14] A. Ceriello, L. Monnier, and D. Owens, "Glycaemic variability in diabetes: Clinical and therapeutic implications," *Lancet Diabetes Endocrinol.*, vol. 7, no. 3, pp. 221–230, 2019, doi: 10.1016/S2213-8587(18)30136-0.

[15] J. Daniels, P. Herrero, and P. Georgiou, "Personalised glucose prediction via deep multitask networks," in *Proc. 5th Int. Workshop Knowl. Discov. Healthcare Data Co-Located with 24th Eur. Conf. Artif. Intell.*, ser. CEUR Workshop Proceedings, vol. 2675. K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., CEUR-WS.org, 2020, pp. 110–114. [Online]. Available: http://ceur-ws.org/Vol-2675/paper19.pdf

[16] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," in *Proc. IEEE*, vol. 109, no. 1, Jan. 2021, pp. 43–76. doi: 10.1109/JPROC.2020.3004555.

[17] W. Zhang *et al.*, "Deep model based transfer and multi-task learning for biological image analysis," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*. Sydney, NSW, Australia: ACM Press, 2015, pp. 1475–1484. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2783258.2783304

[18] T. Killian, G. Konidaris, and F. Doshi-Velez, "Transfer learning across patient variations with hidden parameter Markov decision processes," 2016, *arXiv:1612.00475*.

[19] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 200–213, Apr.–Jun. 2020.

[20] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," 2019, *arXiv:1812.11806*.

[21] A. R. Bhimireddy, P. Sinha, B. Oluwalade, J. W. Gichoya, and S. Purkayastha, "Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks," in *Proc. 5th Int. Workshop Knowl. Discov. Healthcare Data, Eur. Conf. Artif. Intell.,* ser. CEUR Workshop Proceedings, vol. 2675, K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., CEUR-WS.org, 2020, pp. 125–130. [Online]. Available: http://ceur-ws.org/Vol-2675/paper22.pdf

[22] H. Rubin-Falcone, I. Fox, and J. Wiens, "Deep residual time-series forecasting: Application to blood glucose prediction," in *Proc. 5th Int. Workshop Knowl. Discov. Healthcare Data Co-Located with 24th Eur. Conf. Artif. Intell.,* ser. CEUR Workshop Proceedings, vol. 2675, K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., CEUR-WS.org, 2020, pp. 105–109. [Online]. Available: http://ceur-ws.org/Vol-2675/paper18.pdf

[23] H. Hameed and S. Kleinberg, "Investigating potentials and pitfalls of knowledge distillation across datasets for blood glucose forecasting," in *Proc. 5th Int. Workshop Knowl. Discov. Healthcare Data Co-Located with 24th Eur. Conf. Artif. Intell.*, ser. CEUR Workshop Proceedings, vol. 2675, K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., CEUR-WS.org, 2020, pp. 85–89. [Online]. Available: http://ceur-ws.org/Vol-2675/paper14.pdf

[24] W. Gu, Z. Zhou, Y. Zhou, M. He, H. Zou, and L. Zhang, "Predicting blood glucose dynamics with multi-time-series deep learning," in *Proc. 15th ACM Conf. Embedded Netw. Sensor Syst. - SenSys*. Delft, The Netherlands: ACM Press, 2017, pp. 1–2. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3131672.3136965

[25] M. De Bois, M. A. E. Yacoubi, and M. Ammi, "Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people," *Comput. Methods Programs Biomed.*, vol. 199, 2021, doi: 10.1016/j.cmpb.2020.105874.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," Mar. 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[27] T. Kushner, M. D. Breton, and S. Sankaranarayanan, "Multi-hour blood glucose prediction in type 1 diabetes: A patient-specific approach using shallow neural network models," *Diabetes Technol. Therapeutics*, vol. 22, no. 12, pp. 883–891, 2020. [Online]. Available: https://www.liebertpub.com/doi/full/10.1089/dia.2020.0061

[28] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.

[29] C. Marling and R. C. Bunescu, "The ohiot1dm dataset for blood glucose level prediction: Update 2020," in *Proc. 5th Int. Workshop Knowl. Discov. Healthcare Data Co-Located with 24th Eur. Conf. Artif. Intell*. ser. CEUR Workshop Proceedings , vol. 2675, K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, Eds., CEUR-WS.org, 2020, pp. 71–74. [Online]. Available: http://ceur-ws.org/Vol-2675/paper11.pdf

[30] L. Monnier *et al.*, "Toward defining the threshold between low and high glucose variability in diabetes," *Diabetes Care*, vol. 40, no. 7, pp. 832–838, Jul. 2017.

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

[32] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," in *Proc. AMIA Summits Transl. Sci.*, 2020, pp. 191–200.

[33] E. I. Georga *et al.*, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 1, pp. 71–81, Jan. 2013.

[34] M. Gadaleta, A. Facchinetti, E. Grisan, and M. Rossi, "Prediction of adverse glycemic events from continuous glucose monitoring signal," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 2, pp. 650–659, Mar. 2019.

[35] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[36] F. Chollet *et al.*, "Keras: The Python deep learning library," Astrophysics Source Code Library, 2018.

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

[38] W. L. Clarke, "The original clarke error grid analysis (EGA)," *Diabetes Technol. Therapeutics*, vol. 7, no. 5, pp. 776–779, Oct. 2005.

[39] G. P. Forlenza *et al.*, "Predictive low-glucose suspend reduces hypoglycemia in adults, adolescents, and children with type 1 diabetes in an at-home randomized crossover study: Results of the PROLOG trial," *Diabetes Care*, vol. 41, no. 10, pp. 2155–2161, Oct. 2018.

[40] W. A. Pruett, J. S. Clemmer, and R. L. Hester, "Physiological modeling and simulation-validation, credibility, and application," *Annu. Rev. Biomed. Eng.*, vol. 22, no. 1, pp. 185–206, 2020

[41] Comet.ML, "Comet supercharge machine learning," Accessed: Feb. 3, 2021. [Online]. Available: https://www.comet.ml/.