Data-driven Glucose Prediction for Type 1 Diabetes

Modeling, Adaptation and Engineering

Ran Cui

A thesis submitted for the degree of Doctor of Philosophy

School of Computing
ANU College of Engineering and Computer Science

2 March 2025

Declaration

This dissertation is an account of research undertaken between July 2019 and August 2024 at the School of Computing, College of Engineering, Computing & Cybernetics, The Australian National University, Canberra, Australia. The work presented in this thesis is that of the candidate alone, except where indicated by due literature reference and acknowledgements in the text. It has not been submitted in whole or in part for any other degree at this or any other university.

Ran Cui
2 March 2025

Acknowledgements

This is indeed a unique and irreplicable journey. I would like to express my gratitude to those who have shared their memory with me along the way.

My deepest thank you goes to my supervisor Prof. Hanna Suominen, for the countless support and dedication she offered, the meticulous help she committed in reviewing my work, and the attempts we made together in exploring various research possibilities; and Dr. Elena Daskalaki, for her invaluable inspiration, the productive discussions, and the unwavering help even after her formal commitment with the university. Special thanks to my co-supervisors: Prof. Christopher Nolan, for his insightful advice and enduring patience; and Dr. Md Zakir Hossain, for the valuable discussions.

I am also grateful to work together and get along with colleagues Prof. Inger Mewburn, Dr. Lizhen Qu, Dr. Qiongkai Xu, Dr. Artem Lenskiy, and Dr. Pai Peng. The PhD study has been generously supported and funded by the Our Health In Our Hands grand challenge, School of Computing of the Australian National University, and National Computational Infrastructure Australia. The internship midway through the PhD study has been supported by Bilibili Inc. I thank them with great gratitude for all.

Life in Canberra is an unfading memory. I have had the best fortune to develop the close bond with Teng Ma, Binyuan (Amber) Chu, He Diao, Lingyu Xia, and Wanlin Wu. Through the years, I share precious memory with devoted friends Haidi Hong, Zhuoran Liu, Yan Li, Yumo Dong, Xiaoxue Zhao, Xiaoxuan Du, Robin Vlieger, Chirath Hettiarachchi, Sandaru Seneviratne, Chenchen Xu, Dongxu Li, Hua Hua, Weixuan Sun, Nicholas I-Hsien Kuo, Wenbo Ge, Li'An Chen, and many more. Thank you all for making time so colorful.

During the COVID-19 pandemic I took part of the study remotely in Shanghai. I am fortunate enough to share the sparkling moments with Yaxin (Albert) Wang, Zimo Zhu, Hongkuan Yu, Ruozhu Li, Charles Park, Xiaojing (hyo) Huang, Yuehua Li, Xuzheng Fu, Jingjing (Krystal) Ge, Xuan Zhao, Yining (Connie) Tang, Tianwen Qian, Ruodan Yan, Jingtai Zheng, Ao Feng, Wanxin Yao, and Xudong Zhang.

Finally, I would like to thank my family for all the love and supports, without which this accomplishment would never be possible. I have been away from home for too long, and it is time to get back to be around you.

Abstract

Type 1 Diabetes Mellitus (T1DM) is a chronic disease that affects over 9 million people worldwide, people with T1DM require continuous external insulin delivery to control blood glucose levels. Over the years, the development of Continuous Glucose Monitoring (CGM) systems has enabled people with T1DM to constantly monitor the glucose trajectory. The availability of CGM data has also led data-driven glucose prediction using Machine Learning (ML) to a prominent focus in current biomedical research.

The aim of this PhD thesis in computer science is to advance data-driven glucose prediction research through three objectives: modeling, adaptation and engineering. More specifically, the modeling objective is to enhance the existing modeling techniques used in data-driven glucose prediction, e.g., the Recurrent Neural Network (RNN) based modeling and the Convolutional Neural Network (CNN) based modeling. The adaptation objective focuses on the specific problem of adapting generic glucose prediction models to an individual patient. This would be especially useful at the onset of using personalised glucose prediction, in which the volume of personalised data is limited. The engineering objective addresses the practical implementation of data-driven glucose prediction. This involves optimising the computational efficiency of the models, improving data processing and training & testing instances extraction, and providing suggestions on different implementation options. Driven by the three objectives, this thesis reports on five individual studies undertaken during my PhD research.

In the first study, I studied the correlation between the presence of diabetes and a series of blood biomarkers including blood glucose. In investigating 9,549 people in the China Health and Nutrition Survey dataset, I formulated the ML classification task of using the biomarkers together with the age and gender information of each person to classify if they were with diabetes, and the employed algorithm XGBoost achieved over 86% F1 score in the experiments. A features importance test was subsequently performed on the trained model. It was validated that the blood glucose was dominant in determining the presence of diabetes and no other blood biomarkers were more important than age. The results demonstrated the significance of investigating glucose as a biomarker for diabetes, and encouraged the following research on data-driven glucose prediction, which is the main focus of this thesis.

In the second study on modeling and adaptation, I studied the use of a self-attention mechanism in glucose prediction. Building on previous work using RNNs and Temporal Convolutional Networks (TCN), for the first time I proposed using the self-attention mechanism as the primary modeling technique, recognising its superior learning capacity. More specifically, I proposed the Recurrent Self-Attention Network (RSAN) which predicted multi-step glucose by unrolling the prediction one at a time. In the experiments on the OhioT1DM dataset, RSAN achieved state-of-the-art predictive performance of root mean squared error 17.82 mg/dL for 30 minutes and 28.54 mg/dL for 60 minutes with statistical significance compared to a series of benchmarks at the time. Moreover, transfer learning was applied to perform the instance adaptation in which an RSAN model was pre-trained using data from non-target patients and then fine-tuned using the target patient's personal data. The use of transfer learning was demonstrated to contribute around 1 mg/dL improvement. The results demonstrated the validity of the proposed RSAN model in glucose prediction and the use of transfer learning with it.

In the third study that comprehensively covered modeling, adaptation, and engineering, I moved to a new direction of time-index based modeling, which was in contrast to existing studies using historical-value modeling. More specifically, I evaluated using meta-optimised time-index model (MOTIM) as a new methodology in glucose prediction and its instance adaptation. Instead of learning a mapping from the recent glucose sequence to predict future glucose as in historical-value modeling, the time-index model learned a mapping from the time-index feature to the corresponding glucose value, and the use of meta-optimisation enabled the time-index model to adjust to the local glucose pattern. As the learning objective reduced from a sequence-to-sequence mapping to a point-to-point mapping, the efficiency of the algorithm was largely improved. MOTIM achieved performance comparable to state-of-the-art historical-value models in the personalised training experiment and the instance adaptation experiment on the OhioT1DM dataset and UMT1DM dataset. For example, it achieved 1-hour excursion averaged mean absolute percentage error of 10.24 on OhioT1DM and 12.57 on UMT1DM in the personalised training experiment which were similar to existing historical-value methods, while consuming a substantially lower cost, e.g., only 1.3% model parameters and more than 10 times faster inference. These promising findings demonstrated the validness of the proposed methodology, and would encourage further work on using time-index modeling for data-driven glucose prediction.

In the fourth study on engineering, I focused on the problem of postprandial glucose prediction. Different from prior research which focused on hypoglycemia only, I made the first attempt to establish the problem definition as the joint prediction of hyperglycemia and hypoglycemia. Instead of training two models for the two prediction targets, I unified the two targets into one Long Short-Term Memory based backbone with two linear heads for the two targets, thus improving the engineering efficiency. The evaluation of this methodology on the OhioT1DM dataset achieved Matthew's correlation coefficient of 0.61 for hyperglycemia and 0.48 for hypoglycemia, outperforming a variety of studies in generic glucose prediction, postprandial glucose prediction and a series of traditional baselines that I implemented. These findings demonstrated the validity of the proposed joint problem definition of predicting postprandial hyperglycemia and hypoglycemia.

In the last study on engineering, I targeted comparing the different engineering practices observed in existing research. More specifically, existing research has used rolling prediction and direct prediction as the prediction schemes, and the single-point prediction and excursion prediction as the prediction targets. Due to the necessity for ML practitioners to make exclusive decisions about these engineering options, I conducted comparative experiments to analyse their effects on predictive performance and efficiency. The results using the OhioT1DM dataset clearly demonstrated the better performance of direction prediction over rolling prediction, and the excursion prediction over the single-point prediction. These findings would not only enhance predictive accuracy of the models but also improve the efficiency and practicality of the ML pipeline.

The contributions made by this thesis to modeling, adaptation, and engineering of data-driven glucose prediction could enlighten the future research towards precise and reliable glucose prediction technology for people with T1DM. To encourage such progress, open-source code releases have been made available to supplement this thesis.

Keywords: deep learning, evaluation study, glucose prediction, machine learning, medical informatics, type 1 diabetes, validation study

Contents

Figures

Tables

Key Abbreviations

ANU	**A**ustralian **N**ational **U**niversity
CGM	**C**ontinuous **G**lucose **M**onitoring
DL	**D**eep **L**earning
FPG	**F**asting **P**lasma **G**lucose
GRU	**G**ated **R**ecurrent **U**nit
HbA1c	**H**emoglobin **A1c**
LSTM	**L**ong **S**hort-**T**erm **M**emory
MAC	**M**ultiply-**AC**cumulate
MCC	**M**atthew's **C**orrelation **C**oefficient
ML	**M**achine **L**earning
MLP	**M**ulti**l**ayer **P**erceptron
MOTIM	**M**eta-**O**ptimised **T**ime-**I**ndex **M**odel
RMSE	**R**oot **M**ean **S**quare **E**rror
RNN	**R**ecurrent **N**eural **N**etwork
RSAN	**R**ecurrent **S**elf-**A**ttention **N**etwork
TCN	**T**emporal **C**onvolutional **N**etwork
T1DM	**T**ype **1** **D**iabetes **M**ellitus
T2DM	**T**ype **2** **D**iabetes **M**ellitus

Research Outputs

Publications

The following is the list of peer-reviewed publications.

R. Cui, E. Daskalaki, M. Z. Hossain, A. Lenskiy, C. J. Nolan and H. Suominen (2021). 'A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning'. In: *2021 15th International Congress in Nursing Informatics (NI)*. IOS Press, pp. 36–38. DOI: 10.3233/SHTI210657.

R. Cui, C. Hettiarachchi, C. J. Nolan, E. Daskalaki and H. Suominen (2021). 'Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network'. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. DOI: 10.1109/CBMS52027.2021.00064.

R. Cui, C. J. Nolan, E. Daskalaki and H. Suominen (2023). 'Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes'. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. DOI: 10.1109/EMBC40787.2023.10340094.

The following is a study accepted by peer review and currently in-press.

R. Cui, H. Suominen, C. J. Nolan and E. Daskalaki (2024). 'Meta-optimised Time-index Modeling for Glucose Excursion Prediction'. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 1911–1918.

The following is a study currently under peer review.

R. Cui, E. Daskalaki, C. J. Nolan and H. Suominen (2024). 'Evaluation of Different Engineering Options for Data-driven Glucose Prediction'. In: *TBD*.

The above studies constitute the main body of this thesis. Moreover, the following is the list of publications during my internship at Bilibili Inc. China from July 2021 to July 2022.

R. Cui, T. Qian, P. Peng, E. Daskalaki, J. Chen, X. Guo, H. Sun and Y.-G. Jiang (2022). 'Video moment retrieval from text queries via single frame annotation'. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1033–1043. DOI: 10.1145/3477495.3532078.

T. Qian, R. Cui, J. Chen, P. Peng, X. Guo and Y.-G. Jiang (2023). 'Locate before answering: Answer guided question localization for video question answering'. *IEEE Transactions on Multimedia*. DOI: 10.1109/TMM.2023.3323878.

The following is the research finished before the formal commencement of my PhD study, which has given me my initial motivation to research.

Q. Xu, L. Qu, C. Xu and R. Cui (2019). 'Privacy-aware text rewriting'. In: *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 247–257. DOI: 10.18653/v1/W19-8633.

Presentations

- "A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning", in NI2021 (poster).
- "Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network", in CBMS2021 (oral).
- "Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes", in EMBC2023 (oral).
- "ViGA: Video moment retrieval via Glance Annotation", in SIGIR2022 (oral).

Open-source Releases

- "Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network", https://github.com/r-cui/GluPred, MIT license.
- "Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes", https://github.com/r-cui/PostprandialHyperHypoPrediction, MIT license.
- "Meta-optimised Time-index Modeling for Glucose Excursion Prediction", https://github.com/r-cui/MOTIMGluPred, MIT license.
- "ViGA: Video moment retrieval via Glance Annotation", https://github.com/r-cui/ViGA, MIT license.
- "Evaluation of Different Engineering Options for Data-driven Glucose Prediction", https://github.com/r-cui/EngineeringGluPred, MIT license.

To my family.

Chapter 1

Introduction

Diabetes is a group of metabolic diseases characterised by hyperglycemia resulting from defects in insulin secretion, insulin action, or both (American Diabetes Association, 2010). According to the global disease burden analysis[1], the number of people with diabetes worldwide rose from 108 million in 1980 to 422 million in 2014. In 2019, diabetes was the 9TH most common cause of death in the world (World Health Organization, 2020): around 1.5 million deaths were caused by diabetes, and 48% of all deaths due to diabetes occurred before the age of 70 years (World Health Organization, 2023). It is apparent that diabetes has become a major health burden for humanity.

Diabetes is primarily classified into Type 1 Diabetes Mellitus (T1DM), Type 2 Diabetes Mellitus (T2DM) and others like gestational diabetes (American Diabetes Association, 2010). They exhibit similar symptoms of high blood glucose levels and complications, but have different pathologies. T1DM is caused by autoimmune destruction of pancreatic islet beta-cells, which results in a complete inability to secrete insulin. Thus, it is caused by insulin deficiency. In contrast, T2DM is caused byt insulin resistance and a failure of islet beta-cells to compensate for the insulin resistance with increased insulin secretion. Autoimmune destruction of beta-cells does not occur. As the cause of T1DM is a more absolute insulin deficiency than T2DM, T1DM is dependent on exogenous insulin delivery of to achieve glucose control. The challenge for T1DM patients is to give enough exogenous insulin to prevent hyperglycemia (elevated blood glucose), while avoiding giving too much insulin which can cause hypoglycemia (abnormally low blood glucose). In recent decades, the development of Continuous Glucose Monitoring (CGM) devices (Vettoretti et al., 2020) for T1DM has enabled real-time display of glucose levels at intervals ranging from 1 to 15 minutes. These devices, due to their accuracy, being reasonably unobtrusive, small, comfortable, and user-friendly, hold great potential in informing, educating, motivating, and alerting people with diabetes (Rodbard, 2016). In 2022, 31% of individuals with diagnosed diabetes in America have benefited from the use of CGM devices in their daily lives (American Diabetes Association, 2023). Since CGM devices received first approval for clinical use by the United States Food and Drug Administration in 1999, it is now playing an increasingly important role in diabetes management worldwide.

With the advent of CGM technology providing continuous and detailed glucose data, the integration of advanced analytical methods has become increasingly feasible. Machine Learning (ML) (Jordan and Mitchell, 2015) is the field of building intelligent models by learning from large data, which is capable of generalising to unseen data and thus performing tasks without explicit instructions. A typical example is the handwritten digit recognition (LeCun, Boser et al., 1989). By training on 7,291 pictures of handwritten digits, the model

[1] https://vizhub.healthdata.org/gbd-results/

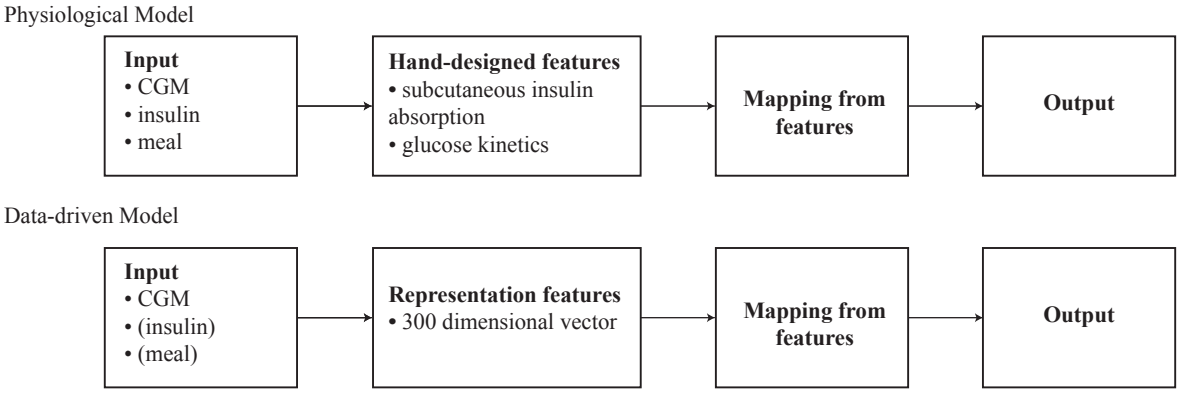


Figure 1.1: Comparing the concepts between a physiological model and a data-driven model. Figure is inspired by the Figure 1.5 in Goodfellow et al. (2016), while adapted to the glucose prediction task. Abbreviations: CGM, Continuous Glucose Monitoring.

is able to achieve an error rate of only 3.4% in recognising unseen pictures of the same kind. Within the broad field of ML, Deep Learning (DL) (LeCun, Bengio et al., 2015), a more recent advancement since the last decade, focuses on tackling more complex tasks and larger volumes of data using models with multiple layers in the model structure and very large trainable parameters. Nowadays, DL plays an essential role in most cutting-edge Artificial Intelligence (AI) applications such as large language models, autonomous driving, speech recognition, and various types of generative AI like image generation and music generation. The increasing feasibility of collecting glucose data through CGM devices further enhances the potential of leveraging ML and DL in diabetes research, opening new avenues for advanced predictive models and personalised healthcare solutions. Data-driven glucose prediction, the research field in which this thesis lies, will be introduced next.

1.1 Data Driven Glucose Prediction

Data-driven glucose prediction (Felizardo et al., 2021) refers to the field in which glucose is predicted using ML techniques on large CGM data. This differs from another traditional methodology for glucose prediction called the physiological model. The physiological model is based on explicit modeling of the metabolism of glucose production, glucose utilisation, and insulin and meal absorption (Oviedo, Vehí et al., 2017). For example, the Hovorka model (Hovorka et al., 2004) is a widely applied physiological model that simulates a series of postprandial metabolism processes, including insulin absorption, carbohydrate digestion, insulin action and glucose kinetics (Calm, García-Jaramillo et al., 2011). Each module in the Hovorka model uses a system of ordinary differential equations parameterised by physiological parameters that need to be set in advance. Those parameters can be estimated through statistical inference methods or simply set to population values (Oviedo, Vehí et al., 2017).

> *The success of ML algorithms generally depends on data representation, and I hypothesize that this is because different representations can entangle and hide more or less the different explanatory factors of variation behind the data* — Bengio, Courville et al. (2013).

This quote, in contrast, reveals the underlying logic of data-driven glucose prediction or any ML-based application, i.e., they attempt to uncover the inherent rules purely through mining the glucose data. To put it differently, the physiological model can be analogous to the classic ML in the early era, in which the focus is to hand design features and apply them to a

relatively simple mapping to the output: physiological glucose prediction models use hand-designed rules that simulate each compartment of the metabolism process. And because of this nature, physiological models are often restricted to specific application scenarios, e.g., the Hovorka model is designed specifically for the fasting condition (Hovorka et al., 2004). Just like ML has evolved from this paradigm to representation learning which is usually tackled by DL, glucose prediction research has also transited to data-driven glucose prediction, where the learning target is to let the model learn more complex representations that cannot be captured by hand design (Goodfellow et al., 2016). As illustrated in Figure 1.1, the key of the data-driven model is that it fully relies on learning from rich CGM data and occasionally with extra physiological inputs, without explicitly describing any physiological process. A representative idea in this field (Allam et al., 2011; Martinsson, Schliep, Eliasson and Mogren, 2020; Zhu, Li, Chen et al., 2020) is to first encode the recent sequence of CGM input $\mathbf{g} = [g_1, g_2, g_3, ...]$ via a Recurrent Neural Network (RNN) (Rumelhart and McClelland, 1987) such as a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) into a deep representation $\mathbf{h}$, then map this representation to the output $g \in \mathbb{R}$ using a simpler learnable mapping such as a linear mapping. Here, the representation $\mathbf{h}$ is relatively more complex, e.g., a 300 dimensional vector. This approach of first encoding the recent glucose sequence into a complex representation and then making prediction is referred to as historical-value modeling in this thesis.

The modern data-driven glucose prediction is advantageous in many aspects. A main benefit is that it does not rely on human expertise, which is potentially inadequate and subjective. The physiological model is based on existing human knowledge of diabetes, while data-driven models leverage the power of ML to reveal the complex patterns and relationships within the data that might be not yet understood by human experts. Another advantage enjoyed by a data-driven model is its versatility — the absence of hand-designed feature makes the learning more flexible. For example, data-driven prediction is not necessarily limited to the fasting condition like in the Hovorka model (Hovorka et al., 2004), as long as such restriction does not exist in the training phase. To accommodate specific conditions, such as breaking the fasting restriction, a physiological model would need to be re-designed, while a data-driven model can stick to its structure to a great extent, and simply change the training data. Moreover, data-driven glucose prediction is capable of continuously improving and adapting as more data become available. Unlike physiological models, which rely on predefined equations and parameters, data-driven models can leverage ML techniques to update and refine their predictions based on new data. This ongoing learning process allows for the integration of diverse data sources, such as meal intake, physical activity, and stress levels, which can all influence glucose levels. Furthermore, data-driven models can be personalised to individual patients (Zhu, Li, Herrero and Georgiou, 2022). By training on personal CGM data and other health metrics, these models can capture unique physiological responses to various factors, leading to more accurate and personalised glucose predictions. This personalised approach is particularly beneficial for managing T1DM, where tight glucose control is crucial. In summary, data-driven glucose prediction models offer numerous advantages over traditional physiological models, including the theoretical higher upper bound in the performance, greater adaptability, personalisation, and practical usability. This thesis, to this end, focuses on improving data-driven glucose prediction from three objectives, namely modeling, adaptation, and engineering, which are to be introduced in the next section.

1.2 Research Objectives

This PhD thesis in computer science, or more specifically, machine learning, data science, and medical informatics, targets the problem of data-driven glucose prediction and aims to improve existing data-driven glucose prediction research through three objectives: **modeling**, **adaptation**, and **engineering**.

The three objectives serve as the central theme throughout this thesis (Figure 1.2). Thus, this section provides an overview of these objectives, outlining the specific areas of focus and the approaches taken to address the challenges in data-driven glucose prediction.

1.2.1 Modeling

The first objective is to enhance the modeling techniques used in data-driven glucose prediction. Current models, predominantly based on RNNs and CNNs, have shown promise but also possess limitations, particularly in handling long-term dependencies and computational efficiency. This thesis has explored novel modeling strategies, including the use of self-attention mechanisms and time-index modeling, to improve predictive performance and efficiency. Chapter 4 and 5 are related to this objective.

1.2.2 Adaptation

The second objective focuses on the adaptation of glucose prediction models to individual patient data. Here, adaptation means instance adaptation, which involves using data from other patients to aid the learning of new patients. The main motivation is that collecting personalised CGM data is often a time-consuming and resource-intensive process, and patients may not have sufficient personal data available at the onset of using these devices. Existing works have investigated using transfer learning and meta-learning approaches for this purpose. This thesis explores these techniques in conjunction with the advanced modeling methods, and introduces innovative approaches to enhance their effectiveness. Chapter 4 and 5 are related to this objective.

1.2.3 Engineering

The third objective concerns the implementation of data-driven glucose prediction. This involves optimising the computational efficiency of the models, improving data processing and training & testing instances extraction, and providing suggestions on different implementation options. This particular objective aims to ensure that the models are not only accurate but also practical for real-world use. Chapter 5, 6, and 7 are related to this objective.

1.3 Research Problems

With the thesis objectives defined, this section provides detailed research questions that this thesis discusses. In the past decades, numerous efforts have been devoted to data-driven glucose prediction targeting many aspects of this problem (Oviedo, Vehí et al., 2017; Felizardo et al., 2021), including novel modeling for learning better representations for glucose prediction, specific predictive models for different scenarios in a day, using non-personalised data to guide the learning of a personalised model, etc. However, there still exists a number of open problems, which this thesis has targeted.

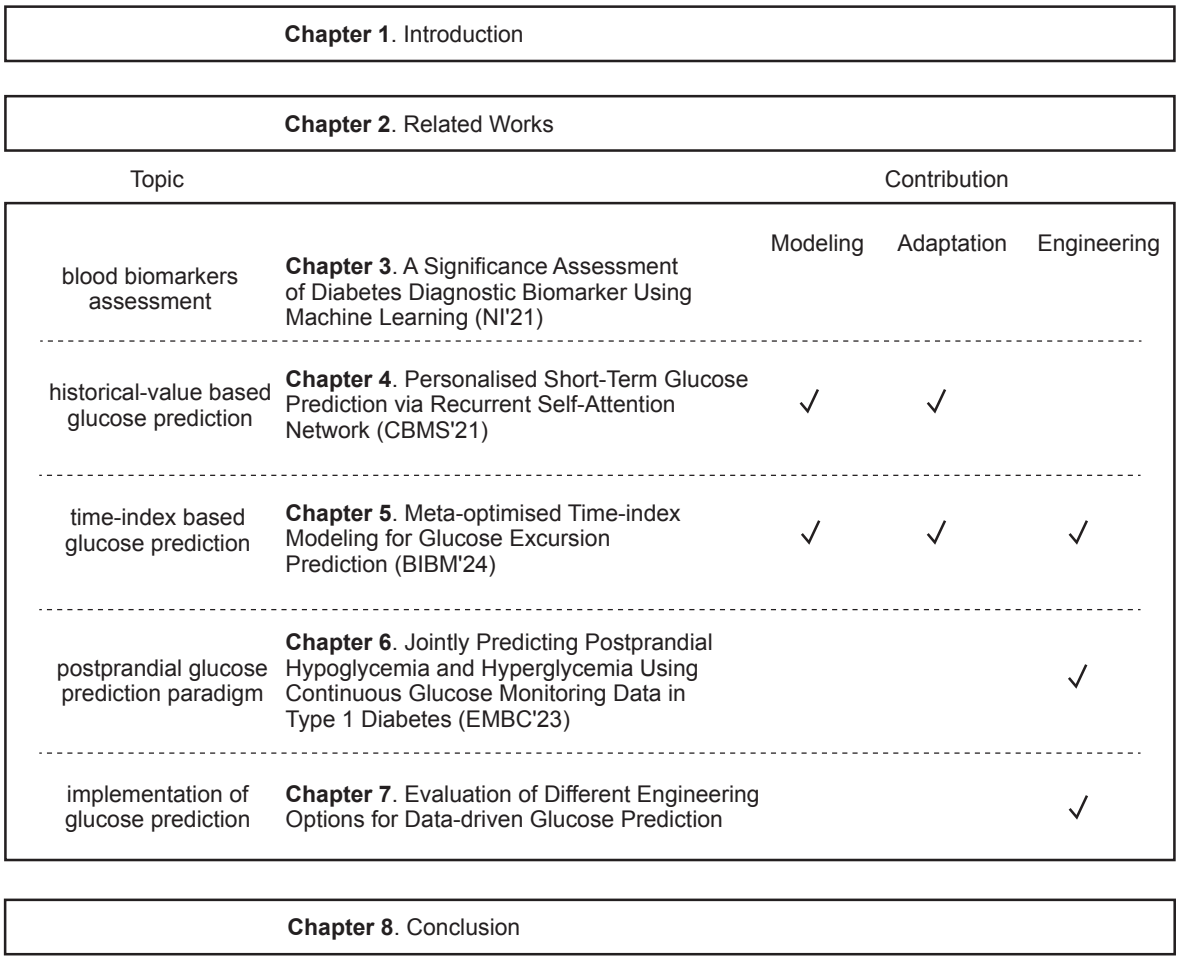


Figure 1.2: Outline of this thesis. Abbreviations: NI, Nursing Informatics; CBMS, International Symposium on Computer-Based Medical Systems; BIBM, International Conference on Bioinformatics and Biomedicine; EMBC, International Conference of the IEEE Engineering in Medicine and Biology Society.

First, as a fundamental aspect of applying ML to specific problems, researchers have explored and compared the performance of various model designs to identify the most effective approaches. Due to the sequential nature of this problem, recent studies are mostly based on recurrent structures as the backbone that encode the input CGM sequence (Allam et al., 2011; De Bois, Yacoubi et al., 2019; Kim et al., 2020; Martinsson, Schliep, Eliasson and Mogren, 2020; Rabby et al., 2021). The first research question investigates whether the modeling can be further improved. Specifically, I explored the potential of novel sequential encoding methods inspired by the advancement of generic ML and DL, such as the self-attention mechanism (Vaswani et al., 2017) and the time-index modeling (Woo et al., 2023), which have gained great success in other applications but have not yet been applied to the problem of data-driven glucose prediction. The discussion related to this research question is presented in Chapter 4 and 5.

Second, I considered the problem of instance adaptation in data-driven glucose prediction. In personalised glucose prediction which aims at obtaining a predictive model that is specific to a single patient, a bottleneck is the lack of personalised data at the beginning of one's access to such facility. To illustrate, in widely applied CGM datasets in research such as the OhioT1DM dataset (Marling and Bunescu, 2020) and the UMT1DM dataset (Fox et al.,

2018), the data per patient have an approximate duration of three months, but it is desirable for a patient to have access to glucose prediction service as soon as possible. Thus, a line of research has been focusing on using data from other patients to guide the learning of a personalised glucose prediction model using approaches of transfer learning (Pan and Yang, 2009; Weiss et al., 2016; Zhuang et al., 2020) and meta-learning (Finn et al., 2017; Vanschoren, 2018; Hospedales et al., 2021). To this end, the second research question investigated whether the improved modeling techniques can be effectively applied to instance adaptation using such techniques and, furthermore, how the concepts of meta-learning can be internalised into the model itself to enhance the effectiveness. Targeting this research question, I evaluated applying transfer learning to the self-attention based model, and the use of meta-optimisation framework with time-index modeling. These investigations are to be presented respectively in Chapter 4 and 5.

Third, I targeted the problem of the postprandial scenario in data-driven glucose prediction. Currently, the research in data-driven glucose prediction is predominantly defined as single-point prediction of a certain prediction horizon, and the evaluation is based on the numeric accuracy such as the Root Mean Squared Error (RMSE). This is a default setup of a generic time-series prediction problem in ML without domain-specific context, so different studies can easily compare to each other. In consideration of the practical application, some studies have moved forward to scenario-specific prediction such as postprandial/non-postprandial (Dubosson et al., 2017; Oviedo, Contreras et al., 2019) or diurnal/nocturnal (Vu et al., 2019; Jensen et al., 2020; Vehí et al., 2020). Nevertheless, such derivative fields are not as mature as the generic glucose prediction, especially those using the data-driven approach. More specifically, studies in this category are not aligned well to each other on experimental conditions and evaluation metrics. For example, (Seo, Lee et al., 2019) studied predicting postprandial hypoglycemia events using daily CGM data only and evaluated the performance using F1 score, yet (Oviedo, Contreras et al., 2019) and (Vehí et al., 2020) studied the same task but only assessed it against a controlled laboratory scenario of a single meal over four hours, and the evaluation, and the evaluation was based on Matthew's correlation coefficient (MCC). These gaps have resulted in a difficulty of cross comparing the performance of the studies to each other. Targeting this problem, I proposed a formalisation of the task of jointly predicting the postprandial hyperglycemia and hypoglycemia events under a pure data-driven manner, defining the procedures of training and testing instances extraction and the corresponding evaluation. This work, categorised as an engineering oriented study, is to be presented in Chapter 6.

Fourth, I targeted the research question of how to improve the efficiency of data-driven glucose prediction models. As an application that mostly needs to run on smart devices, it is important for the algorithm to be sufficiently efficient in terms of time and memory complexity. Since many data-driven models are based on DL which generally consumes large computational resource (Menghani, 2023), managing the algorithm efficiency becomes an important engineering concern. Unfortunately, this is seldom discussed in existing literature of glucose prediction, where the dominant focus still remains on discussing the predictive performance (Oviedo, Vehí et al., 2017; Felizardo et al., 2021). To improve efficiency, rather than focusing solely on coding optimisations, I addressed this research question by enhancing the efficiency through algorithmic improvements and modifications in model design. For example, I investigated the use of a time-index model as a substitution of the widely explored historical-value models, which largely saved the time and memory consumption of the training and inference pipeline. In the study of jointly predicting the postprandial hyperglycemia and hypoglycemia events, I simplified the model to use a shared backbone that predicted the two targets simultaneously, thus the storage resource needed was halved. Related discussions are to be presented in Chapter 5 and 6, respectively.

Fifth, I investigated the inconsistent engineering options found in the existing literature for building ML-based glucose prediction pipelines. It is observed that existing works have adopted different engineering choices for the same module in the implementation of data-driven glucose prediction. For example, for the prediction methodology, some studies have adopted the rolling prediction scheme (De Bois, Yacoubi et al., 2019; Cui, Hettiarachchi et al., 2021; Zaidi et al., 2021) and some adopted direct prediction (Fox et al., 2018; Kim et al., 2020; Rabby et al., 2021; Koca et al., 2023); for the prediction target, existing studies differ in excursion prediction (Calm, Garcia-Jaramillo et al., 2007; Adelberger et al., 2021) or single-point prediction (Dong et al., 2019; Li, Liu et al., 2019; Mirshekarian et al., 2019). These engineering options are mutually exclusive, i.e., one has to select an option and abandon others. In this context, the fifth research question investigates which choice can provide superior performance over its counterpart in terms of predictive accuracy and efficiency. This investigation is to be presented in Chapter 7.

1.4 Thesis Outline

As shown in Figure 1.2, the rest of this thesis is structured as a literature review, five studies, and a conclusion. An overview of the contents of each chapter is as follows.

Chapter 2 reviews the existing works related to this thesis. This chapter is broken down into three sections, each reviewing one objective of this thesis, namely modeling, adaptation, or engineering. As data-driven glucose prediction is a specific application of ML, each review starts with the related developments in generic ML and DL research, then dives into the specific field of data-driven glucose prediction.

Before the detailed discussion on data-driven glucose prediction, Chapter 3 presents a preliminary exploration of using ML as a tool to identify, measure and analyse glucose as a biomarker for diabetes. In this study, a traditional ML model XGBoost (Chen and Guestrin, 2016) was applied to the task of diabetes diagnosis among a Chinese population of 9,549 individuals, using blood glucose together with a series of other blood biomarkers as the input features. With acceptable classification performance achieved, a consequent feature importance assessment indicated that the blood glucose biomarker was dominant in determining the presence of diabetes, while others (albumin, lipoprotein, cholesterol, etc.) were no more important than the age of the individual. This research showed the importance of tracing blood glucose in analysing diabetes and the importance of glucose prediction as the main objective of this thesis.

Chapter 4 presents a novel Recurrent Self-Attention Network (RSAN) for the task of single-point glucose prediction that has been investigated by the majority of data-driven glucose prediction studies (Felizardo et al., 2021). As will be reviewed in Chapter 2, many studies employ recurrent or convolutional structures as the backbone to learn deep representations of the input glucose sequence. However, RNNs suffer from gradient vanishing and exploding problems (Pascanu et al., 2013) which brings difficulty in training, and CNNs are limited in learning the distanced relations in the input sequence. To tackle this, I made use of the self-attention mechanism as the backbone to encode the input glucose sequence, inspired by its effectiveness in other sequential modeling tasks such as natural language processing (Vaswani et al., 2017). The experiments on the OhioT1DM dataset achieved state-of-the-art level performance, hence demonstrating its effectiveness.

Chapter 5 moves the research scope out of historical-value modeling and investigates the use of time-index modeling. In this study, I applied the MOTIM method (Woo et al., 2023) to the task of personalised glucose excursion prediction (El Idrissi, Idri et al., 2020) and

also evaluated the instance adaptation. Different to existing historical-value based methods that learn a sequence-to-sequence mapping, the MOTIM employed time-index modeling that mapped a time-index feature to the glucose value at that time-index. As a result, the MOTIM achieved performance comparable to state-of-the-art historical-value models on the OhioT1DM and UMT1DM datasets, while achieving a substantially lower computing cost, e.g., only 1.3% model parameters and more than 10 times faster inference.

Chapter 6 investigates postprandial glucose prediction, as managing the meal is one of the most important parts in daily glucose management. Existing studies have focused on predicting the postprandial hypoglycemia events, while I argued on the benefits of learning to predict hyperglycemia and hypoglycemia jointly. I formalised this objective by proposing a procedure of extracting traning and testing examples from raw CGM data, and demonstrated the feasibility by using an LSTM backbone with two heads to predict hyperglycemia and hypoglycemia. The experiments on the OhioT1DM dataset demonstrated state-of-the-art performance compared to existing studies on postprandial glucose prediction and several baselines using traditional ML, including Multi-layer Perceptron (MLP), AdaBoost, and Random Forest.

Chapter 7 investigates two engineering modules where inconsistent practices have been observed in existing literature, namely the prediction scheme and prediction target. To achieve this, I built a simple ML pipeline using models that have been proved robust by many of the existing studies in order to best avoid the interference of irrelevant factors, and substituted the targeted engineering practices as a modular component to the ML pipeline. The comparison on predictive performance and efficiency gave clear suggestions that it was preferable to use direct prediction over rolling prediction as the prediction scheme, and excursion prediction over single-point prediction as the prediction target.

The thesis is concluded in Chapter 8 by summarising the principal findings, limitations, and broader impacts. Moreover, Appendix A provides details of the datasets, and Appendix B summarises the implementation details of the studies. Finally, manuscripts of the discussed studies are available in Appendix C for further reference.

Chapter 2

Related Works

This chapter reviews existing works related to the three objectives of this thesis, namely the modeling, adaptation, and engineering of data-driven glucose prediction.[1]

2.1 Modeling

This section reviews the existing studies related to the modeling of data-driven glucose prediction. The review starts with the general development of ML-based time series forecasting methods, then transits to the modeling of data-driven glucose prediction which draws inspiration from generic time series forecasting.

2.1.1 Time Series forecasting

A time series is a sequence of values collected at a defined interval of time. Time series forecasting (Lara-Benítez, Carranza-García and Riquelme, 2021; Benidis et al., 2022; Chen, Ma et al., 2023) is the problem of predicting the future values of a time series based on existing observations. Despite the classical methods such as autoregressive integrated moving average (Box and Pierce, 1970) and exponential smoothing (Gardner Jr, 1985), with the availability of big data and increased computational capacity nowadays, ML-based time series forecasting methods are currently achieving promising results and playing an essential role in research.

To illustrate, a vanilla ML method could be using a linear model to implement the task of time series forecasting as a regular supervised learning problem, i.e., a recent observation of the time series is used as the input, and the desired future values are treated as the output. More formally, to make a prediction of H time steps ahead of a current time t, such a model learns a mapping

$$f([g_t, g_{t-1}, g_{t-2}, ...]) = g_{t+H}.$$

However, there exists a significant drawback in this manner. The values in the input time series sequence are directly used as a collection of unrelated features. In other words, the temporal relation in the time series is disregarded, which substantially weakens the reliability of this approach. Thus, existing research has been focusing on sequence modeling, which aims at capturing the temporal dependencies. Differing in the approach of feature extraction, the models can be classified into RNN-based, CNN-based, and Transformer-based.

[1] Readers are referred to the following survey papers for further reading: Oviedo, Vehí et al. (2017); Zhu, Li, Herrero and Georgiou (2020); Felizardo et al. (2021).

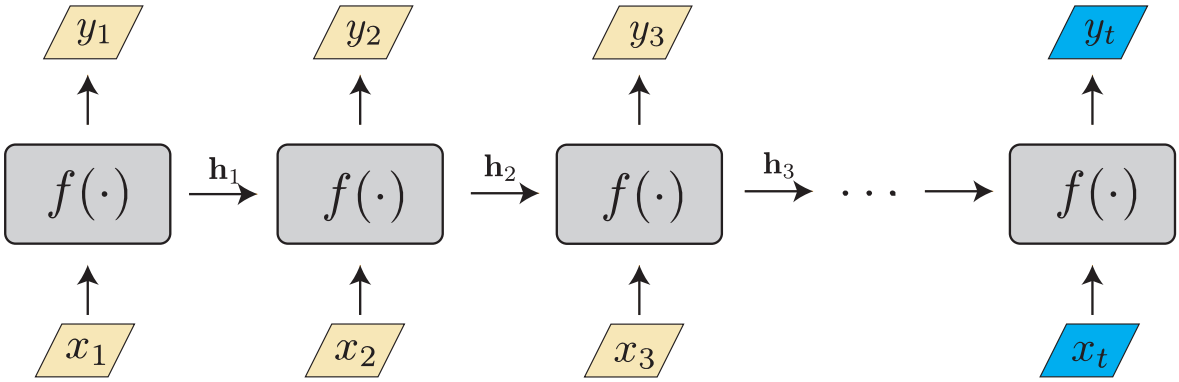
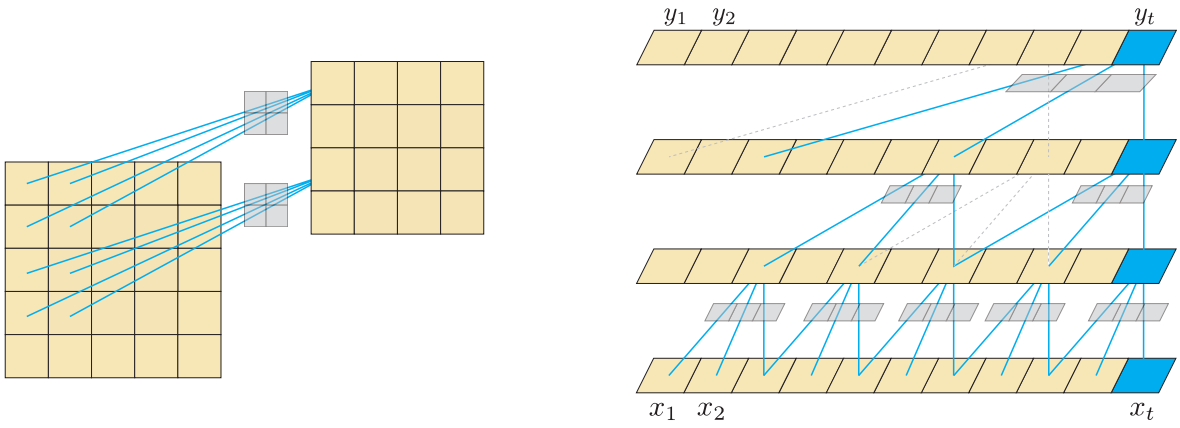


Figure 2.1: Illustration of the feedforward process of a vanilla RNN in handling an input sequence. The same RNN unit recurrently takes a new token x and an inner hidden representation $\mathbf{h}$ of all previously encoded tokens, thus it learns the sequential relation. Abbreviations: RNN, Recurrent Neural Network.

RNN-based model The RNNs, initially investigated by Elman (1990); Jordan (1997), handles the tokens in the input sequence in an autoregressive manner, and holds the capacity of keeping a memory of the previous feedforward process. As illustrated in Figure 2.1, in each step of the RNN feedforward process, the output is determined by the joint input of the current token in the sequence and the hidden state memory from the last step (the first step uses a randomly initialised hidden state). Therefore, the final hidden state is regarded as holding the information of the full sequence. However, a significant limitation of the vanilla RNN is the difficulty of the training (Pascanu et al., 2013). Due to the recurrent nature of RNN, the computation of the gradient, which is the key step of training the model, correlates exponentially to the length of the sequence. To tackle this problem, many variations of the RNN have been proposed, and the most renowned variations are the LSTM (Hochreiter and Schmidhuber, 1997; Graves and Graves, 2012) model and the Gated Recurrent Unit (GRU) (Cho et al., 2014; Chung et al., 2014) model. In the LSTM, the combined use of three gates, respectively called the input gate, the forget gate, and the output gate, enables the LSTM to selectively update the hidden state in each feedforward step. Thus, the vanishing and exploding gradient problem are effectively mitigated. On top of the LSTM, the GRU combines the forget and input gates into a single gate called the update gate, thereby enhancing computational efficiency without compromising performance. The LSTM and the GRU have facilitated a variety of applications that require sequence modeling, such as language modeling, machine translation, and speech recognition. Additionally, the innovation of using the attention mechanisms with RNNs (Bahdanau et al., 2014) has further improved the performance of LSTMs and GRUs by allowing the models to dynamically focus on relevant segments of the input sequence. This enhancement has been particularly beneficial in tasks such as neural machine translation and image description generation.

The research of time series forecasting is substantially influenced by the success of the aforementioned structures, and researchers have been making efforts to refine the RNN models for a better fit to this task using additional techniques such as the attention mechanism (Bahdanau et al., 2014) and the seq-to-seq paradigm (Sutskever et al., 2014). For example, Smyl (2020) combines exponential smoothing and LSTM into a unified framework, in which the exponential smoothing equations enable the method to capture the main components of the individual series, such as seasonality, while the LSTM networks allow non-linear trends and cross-learning. Chang et al. (2017) introduce dilated-RNN, which enables learning the



(a) Convolution operation in 2d image.

(b) Temporal convolutional network.

Figure 2.2: (a) A standard convolution operation on a two-dimensional image, in which the yellow nodes represent pixels in an image and the grey nodes represent a moving kernel. (b) Illustration of the feedforward process of a TCN. Comparing to Figure 2.2(a), the convolution is one-dimensional and causal. Figure 2.2(b) is adapted from Bai et al. (2018). Abbreviations: TCN, Temporal Convolutional Network.

temporal dependencies under different scales. Qin et al. (2017) introduce dual-stage attention-based RNN, in which an input attention and a temporal attention module are respectively applied to the input layer and the hidden feature layer. Du et al. (2020) combine LSTM, attention mechanism, and the seq-to-seq paradigm into a unified model. These advancements illustrate the ongoing evolution and integration of various techniques within RNN models, leading to significant improvements in the accuracy, robustness, and applicability. On top of these efforts, RNN-based models have demonstrated the effectiveness in the problem of time series forecasting across diverse domains, such as predicting climate (Han et al., 2021), power consumption (Nugaliyadde et al., 2019; Shachee et al., 2022), air quality (Athira et al., 2018), etc.

CNN-based model As an alternative method for processing structured data, the CNNs (LeCun, Bottou et al., 1998) has first significantly transformed the field of computer vision. As illustrated in Figure 2.2(a), in one step of the convolution operation, local features are captured via a dot product operation between a moving kernel and snippets at different locations of the input. A complete CNN architecture typically comprises several types of layers, including convolutional, pooling, and fully connected layers, each contributing to the network's ability to learn complex patterns and representations. Pioneering works such as LeNet-5 by LeCun, Bottou et al. (1998) and AlexNet by Krizhevsky et al. (2012) have demonstrated the effectiveness of CNNs in image classification tasks, significantly outperforming state-of-the-art techniques at the time. Subsequent advancements, including the development of deeper networks like VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and Inception (Szegedy et al., 2015), have further enhanced the performance and versatility of CNNs in various computer vision applications, such as object detection, segmentation, and face recognition. The success of CNNs is largely attributed to their capacity to capture local dependencies and their robustness to translation and distortion in images, making them an indispensable tool in modern image analysis and interpretation.

Inspired by the success of CNNs in handling 2D data in computer vision, Oord et al. (2016) adapted the convolution operation to 1D data for the first time, i.e., sequence processing. In their proposed WaveNet model, causal convolutions, as illustrated in Figure 2.2(b), are

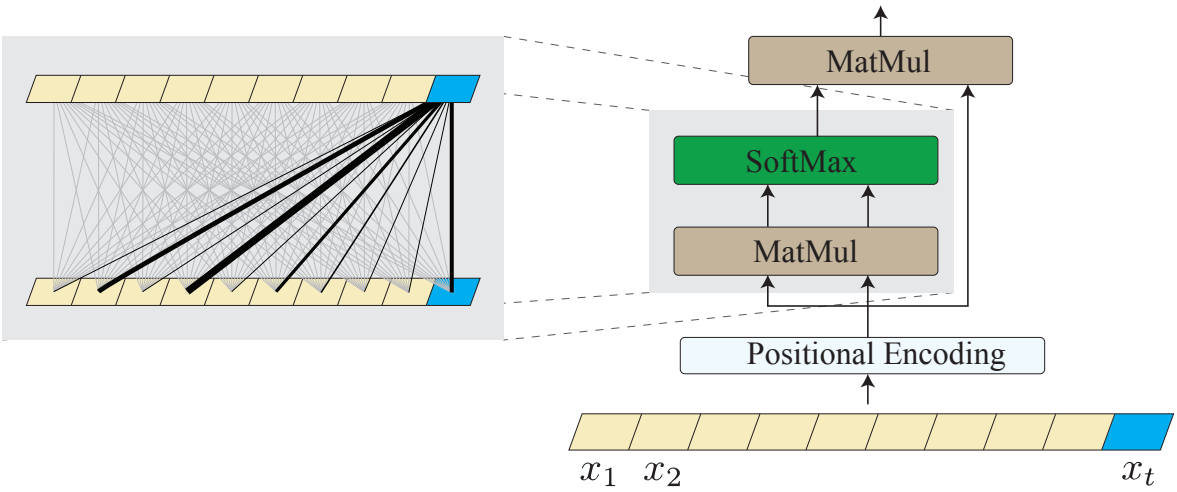


Figure 2.3: Illustration of the self-attention mechanism. The self-attention (called out in the left) is essentially a $t \times t$ weights matrix in which each token in the output is assigned with a weight vector of length t that sums to 1.0. Different thicknesses of the lines indicate that the weight values are different. Figure is adapted from Vaswani et al. (2017); Cui, Hettiarachchi et al. (2021).

used to ensure that the predictions for any given time step depend only on current and past inputs, not future inputs. This design is crucial for generating sequences where the prediction for the next element should not be influenced by future elements, thus preserving the temporal order of the data. This work also incorporates the dilated convolution (Yu and Koltun, 2015) which allows the model to have a larger receptive field without increasing the number of parameters or computational complexity. Following this, Bai et al. (2018) have summarised these contributions into the TCN model and has conducted a comprehensive evaluation of the TCN across a wide range of tasks, including sequential data modeling, audio synthesis, and language modeling. Their results show that TCN is both effective and more stable to train compared to RNN-based models such as LSTM and GRU. Similar to the RNN-based models, TCNs have also demonstrated their value in time series forecasting in various domains, where successful attempts include for example, weather forecasting (Hewage et al., 2020), electricity load (Zhang, Liu et al., 2022), energy demand (Lara-Benítez, Carranza-García, Luna-Romera et al., 2020), and traffic (Zhao et al., 2019).

Transformer-based model Unlike the aforementioned models based on RNN and CNN, the key characteristic of the Transformer (Vaswani et al., 2017) is that it is based solely on attention mechanisms (Bahdanau et al., 2014), and to be more accurate, the self-attention mechanism. As illustrated in Figure 2.3, the self-attention mechanism learns a token-wise weight matrix via a self-product operation, and the weights are then applied to the tokens in the original input sequence. With the augmentation of positional encoding, the self-attention mechanism enables the model to identify the key tokens of interest in the input sequence. The invention of the Transformer model has further revolutionised the DL modeling and is now leading the performance of a significantly wide range of sequence modeling tasks. The early impact of the Transformer is mainly in the field of natural language processing. The BERT model (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) proposes a novel approach of text representation by reading the text bidirectionally, thus allowing a richer context learning. The RoBERTa (Robustly optimised BERT approach) (Liu, Ott et al., 2019) study provides an optimised training regime for BERT, in which the training is with larger mini-batches, more data, and longer sequences. GPTs (Generative Pre-trained Transformer) (Radford et al., 2019; Brown et al., 2020), as the name suggests,

are very large Transformers with billions of parameters that are pre-trained from numerous natural language data. Such large pre-trained Transformers are task-agnostic, meaning no local adaptations towards specific tasks are performed, while they have achieved competitive performance on a wide range of tasks, including machine translation, paraphrasing, question answering, text summarisation, sentiment analysis, and many more.

Moreover, the Transformer has also significantly influenced computer vision research, despite the fact that the image and video data are not sequence data by nature. The first pivotal contribution is the Vision Transformer (ViT) (Dosovitskiy et al., 2020), which has demonstrated that the Transformers could outperform traditional CNNs on image classification tasks when pre-trained on large datasets. The core idea is to segment an image into image patches, thus this high-dimensional data can be represented as a sequence and the requirements of using a Transformer are fulfilled. Following this, a series of successful attempts on different computer vision tasks have been made, including for example, visual understanding (Liu, Lin et al., 2021), object detection (Fang et al., 2021), and segmentation (Ye et al., 2019).

Similar to natural language processing and computer vision, the Transformer model has also significantly influenced the time series forecasting research. The sequential nature of a time series makes it a natural fit to the Transformer concepts. A series of exploration on the original Transformer model has been made towards adapting the model to the time series forecasting task. Li, Jin et al. (2019) propose the LogSparse Transformer that reduces the quadratic complexity of the self-attention mechanism to logarithmic scales. More variations of Transformer are then proposed targeting the long sequence time series forecasting problem, in which the learning of the long temporal dependency and the efficiency of the algorithm are crucial. For example, the Informer introduced by Zhou et al. (2021) has addressed the efficiency and scalability limitations of the original Transformer, and has reduced the quadratic computational complexity by the use of the ProbSparse self-attention mechanism and the self-attention distilling technique. The Autoformer (Wu et al., 2021) has introduced the innovative auto-correlation mechanism as a substitution of the original self-attention. This mechanism identifies sub-series similarities based on the periodicity of the series and aggregates similar sub-series from underlying periods, which effectively reduces the computational complexity. Keeping on this line, the Pyraformer (Liu, Yu et al., 2021) has further introduced the pyramidal attention structure and multi-scale decomposition, which not only captures long-range dependencies, but also achieves time and space complexity linear to the length of sequence. These advancements have significantly broadened the applicability of Transformer models in time series forecasting, especially the long sequence time series forecasting problems such as the exchange rate (Lai, Chang et al., 2018) and electricity transformer temperature (Zhou et al., 2021).

Time-index model The previously reviewed RNN-based, CNN-based, and Transformer-based models, due to the fact that they take a past sequence as the input and map it to the target future values, are referred to as historical-value models in this thesis. Different to historical-values models, time-index model is another category of time series forecasting methods which takes a time-index feature rather than historical values as the input. The Prophet (Taylor and Letham, 2018) is a prestigious time-index model, in which the prediction is decomposed into $y(t) = g(t) + s(t) + h(t) + \epsilon_t$, where $g(t)$, the growth model, models how the time series data have grown and how it is expected to continue growing; $s(t)$, the seasonality model, models the periodic changes such as weekly and yearly seasonality; and $h(t)$, the holiday model, models the effects of predictable events to the time series, such as the effects of holidays to business activities.

This method has been shown especially effective in forecasting commercial time series. In 2017, Godfrey and Gashler (2017) proposed Neural Decomposition (ND), which is an initial attempt to use time-index based neural networks to fit a time series for forecasting. ND is similar to Fourier decomposition, whilst the components in Fourier decomposition are predetermined (sinusoids) and do not change dynamically, but ND allows adaptive decomposition as more data becomes available. This feature allows extrapolation of the time series by providing the future time indices as the input. Woo et al. (2023) continue developing this idea targeting that ND is reminiscent of classical methods by manually specifying periodic and non-periodic activation functions. More specifically, they apply a meta-optimisation framework on an Implicit Neural Representation model (INR) (Sitzmann et al., 2020) to form a deep time-index model that is capable to forecast time series by time-index extrapolation. Their method, referred to as the Meta-Optimised Time-Index Model (MOTIM) in this thesis, is the current cutting-edge method on time series forecasting using time-index based modeling.

2.1.2 Data-driven Glucose Prediction

Similar to their application across various domains, the aforementioned models have also inspired and influenced the task of glucose prediction. Research in this scope is referred to as data-driven glucose prediction (Oviedo, Vehí et al., 2017), as the common ground of such approaches is that they rely fully on learning from real-world CGM data using ML methodologies. When landing from generic time series prediction to data-driven glucose prediction, the problem is specifically marked by the following characteristics.

First, the smaller volume of data. In the realm of data-driven glucose prediction, the volume of available data is significantly smaller compared to other time series forecasting applications. This limitation is primarily due to the nature of CGM data collection, which involves continuous records of glucose levels in a relatively small population of patients over specific periods. Unlike large-scale datasets such as in financial markets or weather forecasting, CGM data are more limited and patient-specific. This necessitates the development of models that can effectively learn and generalise from limited amount of data, often requiring advanced techniques such as data augmentation, transfer learning, or the integration of additional patient-specific contextual information to enhance model performance.

Second, the smaller and more granular forecasting setup, reflected by the short sampling interval and the short prediction horizon. Modern CGM devices typically provide glucose readings at a static time interval ranging from 1 to 15 minutes. This high-frequency data demand models that can process and make accurate predictions over short time frames. Additionally, the prediction horizon in glucose monitoring is usually short, often spanning from just a few minutes to a few hours ahead. This is in contrast to other time series applications where forecasting horizons can extend to days or months. The shorter prediction horizon in glucose prediction is crucial for timely and actionable insights that can help in immediate decision-making, such as insulin dosing, exercise planning or dietary adjustments.

Third, the specificity of the supplementary modalities. In other prediction problems such as stock prediction, the prediction of a target stock could refer to other modalities of time series such as the market index trend or other stocks in the same industry using the heuristic that they tend to share similar patterns. The problem of data-driven glucose prediction holds the similar capacity as life events such as the food intake and insulin treatments affect the glucose trends. However, the sporadic nature of these events differentiates them from the regular, continuous time series typically seen in other domains. This unique characteristic necessitates a specialised modeling that incorporates customised processing of the extra

modalities than a regular multi-model time series modeling to ensure accurate and reliable predictions.

Fourth, the less seasonality in the glucose data. Many long sequence prediction problems such as weather forecasting, energy consumption prediction and retail sales prediction are heavily influenced by seasonal patterns. These patterns can be annual or linked to specific times of the year like holidays, enable the problem to be traced via learning such recurring patterns. However, in the context of glucose prediction, seasonal influences are less significant. While there might be slight periodical patterns in glucose levels due to a regular lifestyle or dietary habit, the primary factors influencing glucose levels are more immediate and individual-specific, especially under the aforementioned granular forecasting setup where the prediction horizon is usually within 2 hours. This unique aspect of reduced seasonality in glucose prediction emphasises the need for approaches that prioritise short-term fluctuations and individual-specific patterns over long-term seasonal cycles.

Except for confidential data, existing studies in the field have mainly employed the following two openly available datasets to accommodate the above characteristics. This thesis extensively utilises the two datasets to develop and evaluate the proposed models for data-driven glucose prediction.[2]

- **OhioT1DM**: The OhioT1DM dataset (Marling and Bunescu, 2020) was created by Ohio University and first released in 2018. In 2020, the dataset was updated with an additional round of data collection, approximately doubling the available information. As a result, the current version of this dataset contains around eight weeks of CGM data, paired with self-reported insulin treatments, meal events, and physiological sensor data from 12 T1DM patients. This dataset has been used extensively in research of data-driven glucose prediction, see for example, Li, Liu et al. (2019); Deng, Lu et al. (2021); Rabby et al. (2021).
- **UMT1DM**: This thesis refers to the dataset created by the US University of Michigan as UMT1DM (Fox et al., 2018) as the author of this dataset did not provide an official name. It contains daily CGM data of 40 T1DM patients under a three-year protocol: every three months, patients are required to wear the CGM device for several consecutive days. Different to OhioT1DM, this dataset contains CGM data only, without any exogenous information. Despite not as extensively used as OhioT1DM, this dataset has also demonstrated its potential for contributing to research in data-driven glucose prediction, see for example, Armandpour et al. (2021).

Numerous studies have focused on leveraging ML algorithms to improve the accuracy and reliability for glucose predictions. As a specific application of generic time series forecasting, research in the broader field has also significantly influenced and advanced data-driven glucose prediction. Similar to generic time series forecasting, RNN is the first structure investigated in the context of data-driven glucose prediction. The ability to handle sequential data and capture temporal dependencies makes the RNN a natural choice for modeling the complex dynamics of glucose levels. Early attempts of using RNN on data-driven glucose prediction can trace back to Mougiakakou et al. (2006). In this study, the performance of MLP and RNN is evaluated and compared in predicting glucose trajectory using data from 4 children with T1DM, and the results suggest that the RNN outperforms a regular MLP. Similarly in the study by Allam et al. (2011), 9 people with T1DM are required to collect their CGM data for 2 days. The evaluation using a MLP and a RNN to predict the glucose value up to 60 minutes indicates that the RNN gives more accurate predictions. Later on, the upgraded versions of RNN, namely LSTM and GRU, have been investigated in glucose

[2]More detailed description of the datasets can be found in Appendix A.

prediction. For example, Martinsson, Schliep, Eliasson and Mogren (2020) have evaluated the performance of the vanilla LSTM model using the OhioT1DM dataset (Marling and Bunescu, 2020). Not only showing the capability of the LSTM model in data-driven glucose prediction, their comparative experiments also suggest that the model trained via using all 6 patients in the OhioT1DM dataset (the 2018 version) has achieved the globally best performance, suggesting the existence of patterns governing blood glucose variability that can generalise between different patients. Kim et al. (2020) have evaluated and compared the original RNN and the two variations LSTM and GRU in predicting the 30-minute glucose value in the future using data from 20 patients with T2DM. The experiments suggest that the GRU model perform slightly better than the rest. This diverse range of studies demonstrates the ongoing refinement and application of RNN-based models in the field of glucose prediction.

Regarding CNN-based models, a pioneering study by Zhu, Li, Herrero, Chen et al. (2018) applied the causal convolution to the task of data-driven glucose prediction. Their model highly replicates the WaveNet structure (Oord et al., 2016) and has achieved performance comparable to the state-of-the-art RNN-based models. Similarly, Bhargav et al. (2021) utilised TCN as a global predictor for multiple simulated patients, and has demonstrated its predictive performance superior to a regular MLP. On top of these explorations on the sole CNN-based structure, Li, Daniels et al. (2019) have proposed the convolutional recurrent neural network which exploits the merits of both CNN and RNN. This model first encodes the raw input glucose sequence into a sequential representation using regular convolution operation (in contrast to causal convolution), then forwards it to an LSTM model. With regard to comparing RNN-based models and CNN-based models, Bhimireddy et al. (2020) have conducted an evaluation that compares LSTM, BiLSTM, convolutional LSTMs, and the TCN, and concludes that the sequence-to-sequence BiLSTM performs better than the other models. However, in the experiments from another study attributed to El Idrissi and Idri (2020), the CNN-based model has achieved better predictive performance than LSTM. This indicates that the effectiveness of different neural network architectures for glucose prediction can vary based on the specific dataset and experimental setup, highlighting the importance of standardised evaluation procedures, which is one of the motivations of my study to be presented in Chapter 6.

More recent advancements of data-driven glucose prediction have increasingly focused on using Transformer or the self-attention mechanism to enhance model accuracy and robustness. The pioneering study is my paper Cui, Hettiarachchi et al. (2021) to be presented in Chapter 4. Following this, Acuna et al. (2023) have compared the performance of Transformer with other models, including XGBoost and 1D-CNN, and has concluded on the better performance of the Transformer. Sergazinov et al. (2023) introduce GluFormer, which utilises the Transformer architecture for personalised glucose forecasting. This model incorporates uncertainty quantification, making it a more robust and interpretable tool. Zhu, Chen et al. (2023) have proposed an innovative edge-based temporal fusion Transformer and has demonstrated its implementation on wearable devices. These studies collectively highlight the potential of self-attention mechanisms in improving the predictive capabilities and practical applications for data-driven glucose prediction.

The above survey has reviewed the current state of data-driven glucose prediction research in terms of the modeling, where existing works have mostly approached the problem via historical-value models. As time-index based modeling is a promising new direction in generic time series forecasting, my research of applying the MOTIM method to data-driven glucose prediction is to be presented in Chapter 5. Besides, the aforementioned work of applying self-attention mechanism as the core modeling backbone is to be presented in Chapter 4.

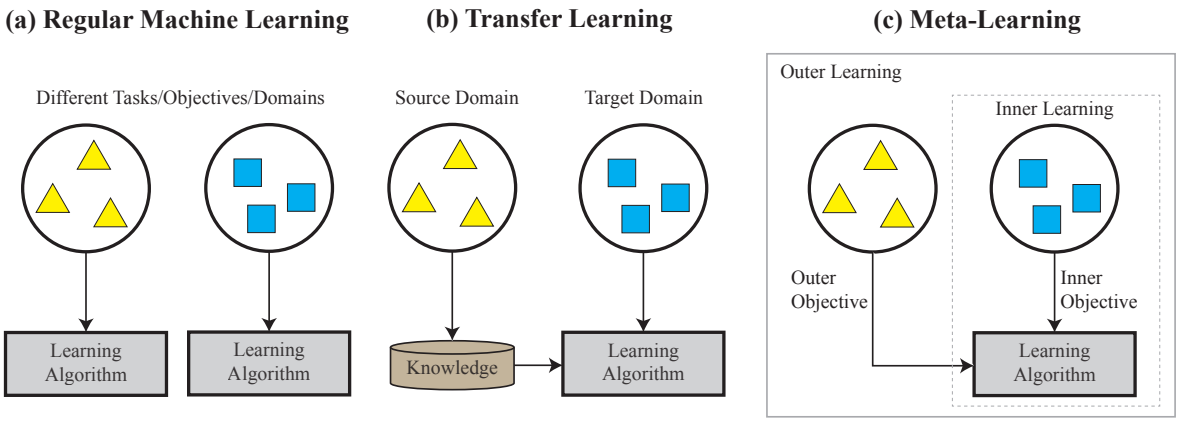


Figure 2.4: Illustrating transfer learning and meta-learning in comparison with standard ML. The concepts are summarised from Zhuang et al. (2020) for transfer learning and Hospedales et al. (2021) for meta-learning. Figure gets inspiration from Pan and Yang (2009).

2.2 Adaptation

This section reviews the existing studies related to the adaptation of data-driven glucose prediction, which involves adjusting predictive models to accommodate the individual variations. The review starts with the general development and application of transfer learning and meta-learning which are the core techniques used for this purpose, then transits to reviewing the state of the art in terms of adaptation in data-driven glucose prediction.

2.2.1 Transfer Learning and Meta-Learning

Existing research on the instance adaptation in data-driven glucose prediction mostly applies meta-learning and transfer learning on the historical-value models to conduct instance adaptation. To this end, this section sorts out the relevant concepts of meta-learning and transfer learning.

In early literature, the concepts of meta-learning and transfer learning can be vague. Many perspectives of meta-learning can be found in the literature as discussed by Vilalta and Drissi (2002), and sometimes the term meta-learning and transfer learning are used interchangeably, such as in Pan and Yang (2009). In contemporary literature (Zhuang et al., 2020; Hospedales et al., 2021), the distinctions between meta-learning and transfer learning have become more clearly defined, and meta-learning is generally regarded as a super set of transfer learning. Meta-learning, also termed as "learning to learn", refers to a broad paradigm and theoretical framework that concerns improving a learning algorithm over multiple learning episodes with different objectives. Transfer learning, on the other hand, refers to the concrete family of solutions for transferring knowledge learned in other domains to a target domain. A core approach in transfer learning is the model parameters transfer, while meta-learning holds a broader umbrella dealing with a much wider range of "meta-representations". The meta-representation in meta-learning is an abstract concept — it can be model parameters initialisation like in transfer learning, or others including optimiser choice, batch selection, hyperparameters and many more, depending on the objective for applying the meta-learning. The difference of meta-learning and transfer learning is presented in Figure 2.4 in the way of comparing to traditional ML. A detailed narration of the relations among meta-learning, transfer learning, and even more relevant concepts including domain adaptation, continual learning, and lifelong learning can be found in Hospedales et al. (2021).

Definition 1. (Meta-Learning) During *base learning*, an *inner* (or lower/base) learning algorithm solves a task such as image classification, defined by a dataset and objective. During *meta-learning*, an *outer* (or upper/meta) algorithm updates the inner learning algorithm such that the model it learns improves an outer objective. For instance this objective could be generalisation performance or learning speed of the inner algorithm — Hospedales et al. (2021).

The above definition reveals the broad scope of meta-learning. This is primarily due to the flexibility in definitions of inner and outer objectives, making meta-learning a more versatile paradigm than transfer learning. In practice, this versatility allows meta-learning to be applied across various domains and tasks, each with unique requirements and challenges. For instance, targeting few-shot learning, meta-learning frameworks like Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) and Prototypical Networks (Snell et al., 2017) have demonstrated remarkable effectiveness in learning a generic initialisation of model through various tasks. By setting the outer objective of the meta-learning to improving generalisation ability of the inner model, these approaches significantly reduce the need for large annotated datasets, which are often expensive and time-consuming to obtain. Similar ideas are also presented in reinforcement learning, a sub-field of ML that focuses on training agents to make sequences of decisions by maximising cumulative rewards through interactions with an environment. Algorithms such as RL2 (Duan et al., 2016) and MetaGenRL (Kirsch et al., 2019) leverage meta-learning principles to enable agents to quickly adapt to new environments or tasks with minimal retraining. In these studies, the meta-representation to be learned is the initial condition. Another important meta-representation is the optimiser. Andrychowicz et al. (2016) have proposed the meta-learning framework whose outer objective is to obtain the best optimiser parameterised by the learning rate, decay coefficients, etc. By leveraging a meta-learner to optimise the learning process itself, this framework opens new possibilities for developing more flexible and robust ML models. More forms of the meta-representation, such as the hyperparameter (Yang and Hospedales, 2016; Franceschi et al., 2018; Lin, Baweja et al., 2019), metric (Koch et al., 2015; Vinyals et al., 2016; Sung et al., 2018), and loss function (Wichrowska et al., 2017; Bechtle et al., 2021) have also demonstrated remarkable success in enhancing the learning effectiveness. These advancements of meta-learning that covers wide range of topics in ML have specifically shown the flexibility of meta-learning, brought by the unlimited possibilities in defining the inner and outer objectives.

Definition 2. (Transfer Learning) Given some observations about the *source domain(s)* and some observations about the *target domain(s)*, transfer learning utilises the knowledge implied in the source domain(s) to improve the performance of the learned decision functions on the target domain(s) — Zhuang et al. (2020).

Inspired by intuitive examples in real life such as one's learned skill of riding a bicycle can help learn to ride a motorcycle, transfer learning aims to leverage knowledge learned from one domain (named the source domain) to aid the learning of another domain (named the target domain). As transfer learning concerns solely transferring the knowledge learned across domains, it is considered as a specific case of meta-learning. That is, when the outer objective of meta-learning is set to aid the inner learning via knowledge learned from the outer learning, the meta-learning concepts become equivalent to transfer learning.

There exists a wide variety of approaches in transfer learning, categorised into instance-based, feature-based, and model-based (Zhuang et al., 2020). Instance-based transfer learning deals with the instance weighting strategy. This technique assigns different weights to the training examples from both domains in training, in which the target domain usually gets higher weights due to its limited number of instances. This simple idea has been shown specifically effective, such as in natural language processing (Jiang and Zhai, 2007) and computer vision

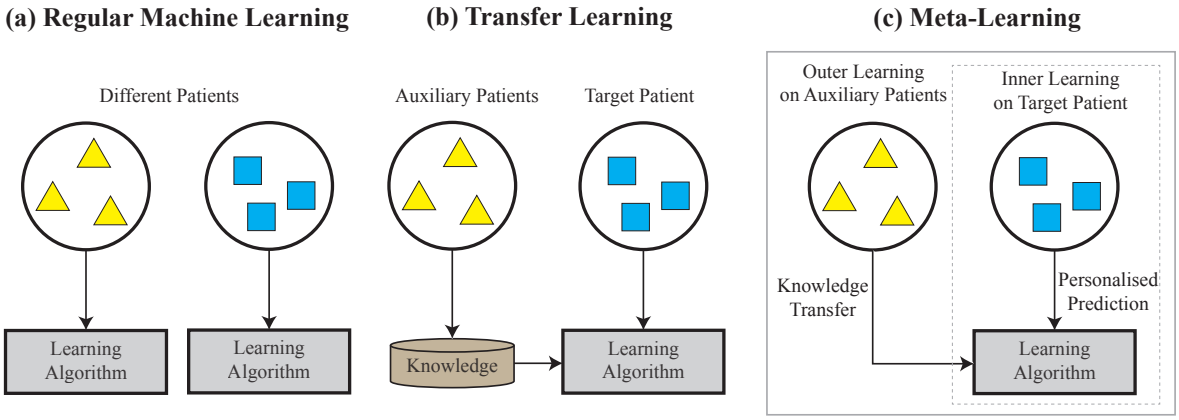


Figure 2.5: Transfer learning framework in the context of data-driven glucose prediction and a meta-learning framework with same scope. This figure is an instantiation of the generic concepts illustrated in Figure 2.4.

(Lin, Goyal et al., 2017; Cui, Jia et al., 2019). For example, Lin, Goyal et al. (2017) propose focal loss, a dynamically scaled cross-entropy loss that emphasises hard examples, effectively acting as an instance weighting scheme to address class imbalance in object detection tasks. Feature-based transfer learning exploits using the feature extracted by large pre-trained models to a new task, i.e., the target domain. For example, the features extracted by the bottom layers in a CNN pre-trained on the ImageNet dataset (Deng, Dong et al., 2009) are shown to be generalisable to multiple visual tasks like image classification, objection detection and scene recognition (Donahue et al., 2014). Similarly, the GloVe word embeddings (Pennington et al., 2014), which contains pre-trained vector representations for 840 billion words, provides a starting point in performing all kinds of natural language processing tasks. Another approach similar to feature-based transfer learning is the pre-training & fine-tuning approach (Hinton et al., 2006), which belongs to model-based transfer learning. The key difference is that the pre-trained model used in feature-based transfer learning is usually frozen, meaning no further training is performed and the extracted features are directly adopted as given, such as using the GloVe embeddings to translate the input sentence into a sequence of vectors as the first step in a natural language processing task. However, as the name suggests, pre-training & fine-tuning involves a fine-tuning stage of the pre-trained model. This process, which usually uses a smaller learning rate, retains useful features learned during the pre-training phase while adjusting the model to meet the specific requirements of the target task. The most representative example of the pre-training & fine-tuning method is the BERT (Devlin et al., 2018). In this study, the bidirectional Transformer model is first pre-trained on the masked language modeling task, then fine-tuned to more specific tasks such as question answering. There are also many other successful attempts following this, such as Lan et al. (2019); Liu, Ott et al. (2019); Yang, Dai et al. (2019). Another model-based transfer learning approach is parameter sharing. The aim of this method is to enhance overall performance across multiple tasks. In this approach, a portion of the model parameters is shared during the learning processes of these tasks. The key characteristic is that the shared parameters are trained on all tasks, while each task retains its own specialised parameters. These transfer learning approaches, including instance-based, feature-based, and model-based methods, have been proven effective in natural language processing and computer vision, and hold potential for other applications.

2.2.2 Adaptation in Data-driven Glucose Prediction

The above advancements have facilitated research in data-driven glucose prediction exploiting transfer learning (Bhimireddy et al., 2020; Freiburghaus et al., 2020; De Bois, El Yacoubi et al., 2021; Deng, Lu et al., 2021; Seo, Park et al., 2021; Xingsan et al., 2021; Yu, Yang et al., 2021; Thomsen et al., 2023) and meta-learning (Zhu, Li, Herrero and Georgiou, 2022; Zhu, Uduku et al., 2022; D'Antoni et al., 2023; Langarica et al., 2023) for the adaptation objective. A principal motivation of this is the lack of data — comparing to data collection in natural language processing and computer vision where data can be automatically collected from the Internet or other resources, collecting real-world CGM data is much more time consuming as only one data point is collected at each sample interval. Figure 2.5 shows this typical instance of applying transfer learning and meta-learning frameworks in the context of data-driven glucose prediction, i.e., using CGM data of non-target patients to aid the personalised learning for a target patient. In this case, the transfer learning and meta-learning objectives are equivalent, and methodologies in the two fields, e.g., the methods reviewed in Section 2.2.1, can be applied.

Transfer learning has been particularly beneficial in addressing this data scarcity by enabling the transfer of knowledge from models trained on larger datasets to those with limited data. As shown in Figure 2.5, the source domain and target domain under this setup corresponds to auxiliary patients and the target patient. Deng, Lu et al. (2021) have performed a comprehensive evaluation on model-based transfer learning, in which different sets of modules in different models are frozen from the parameters updating in the fine-tuning stage. As a result, a CNN-based model with only the MLP module to be trained in fine-tuning is identified as the best combination. Yu, Yang et al. (2021) introduced a combined use of instance-based and model-based transfer learning techniques. Targeting the lack of personalised data, similar instances from the source domain, measured by the dynamic time warping technique (Deng, Chen et al., 2020), are included into the target domain to form an extended target domain. Subsequently, the full model is pre-trained using the source domain data and fine-tuned with the target domain data. In their ablation studies, this transfer learning methodology achieves RMSE improvements of around 1 mg/dL for a 30-minute prediction horizon and 4 mg/dL for a 60-minute horizon, compared with a benchmark using randomised instance-based transfer learning and removing the model-based transfer learning. De Bois, El Yacoubi et al. (2021) propose a novel adversarial multi-source transfer learning framework which uses feature-based transfer learning. In the scope of multi-source transfer learning, data from the source domain have multiple sources, i.e., data from multiple individuals are used as the source domain data. A basic hypothesis of this feature-based transfer learning is that more general feature representation would transfer better to an unknown individual. Thus, the extracted features from the trained model using source domain data undergo an adversarial training phase where a classifier is trained on discriminating the source patient that the feature originally belongs to. This adversarial transfer learning strategy is shown outperforming standard transfer learning for around 2 mg/dL RMSE for predicting 30-minute glucose. These advancements underscore the significant impact of transfer learning techniques in enhancing the accuracy and reliability of glucose prediction models, demonstrating their potential to improve diabetes management through personalised and adaptive approaches.

Meta-learning, on the other hand, has also been investigated for the same objective of tackling the personalised training under a lack of personalised data (Zhu, Li, Herrero and Georgiou, 2022; Zhu, Uduku et al., 2022). The meta-learning framework proposed by Zhu, Li, Herrero and Georgiou (2022) is a representative study in this scope. In their work, the meta-representation to be learned is the model initialisation. On this basis, the outer learning

objective is to optimise the initial model parameters such that the model could quickly adapt to a new patient with minimal gradient updates, while the inner objective is patient-specific and involves fine-tuning the model to minimise loss for a given patient. Under this setup, the MAML algorithm is employed to solve the meta-learning and obtain the model initialisation. Their experiments of comparing adapting the meta-model and a pre-trained model directly using a global dataset of the auxiliary patients demonstrates the improvement in performance of around 2 mg/dL in predicting 30-minute glucose on the OhioT1DM dataset.

Due to the broader scope of meta-learning than transfer learning, the meta-learning framework for data-driven glucose prediction is not limited to the one shown in Figure 2.5. For example, with the same objective of learning model initialisation as the meta-representation, Langarica et al. (2023) considered that the number of patients in a dataset can be insufficient, thus changed the granularity of meta-learning from patient-based to meal-based to enable more training instances. More specifically, they consider the glucose sequence in a certain horizon after a meal as a meal event, and pair two meal events from a same patient to form a task. As data of one patient usually contain multiple meal events, the number of meta-learning tasks can be largely increased. Another reason of selecting the meal event as the unit of the meta-learning framework is that a meal event would capture the most rich and dynamic information within glucose sequence data. They adopt the Almost No Inner Loop (ANIL) algorithm (Raghu et al., 2019), a variation of the MAML algorithm, to solve the meta-learning problem and find the initialisation of the model. The comparison with standard model-based transfer learning demonstrates the faster convergence of their method. Targeting the prediction of abnormal glycemia events, D'Antoni et al. (2023) exploited a layered meta-learning framework, where the base learner consists of three expert systems (implemented by three neural networks) that respectively handle the prediction of hypoglycemia, hyperglycemia, and euglycemia. The meta-learner, on the other hand, learns the features shared by the three expert systems. Thus, the meta-representation to be learned in this work, which is the shared features, is largely different to the aforementioned studies in which the meta-representation is the model initialisation. These studies have demonstrated a wide range of possibilities in applying meta-learning in data-driven glucose prediction, and the potential in achieving superior performance than transfer learning.

The above survey has reviewed the related works of data-driven glucose prediction in terms of adaptation. As the review suggests, there exists great potential of novel approaches in applying meta-learning in data-driven glucose prediction. Chapter 5 will present the study of how the meta-learning concepts can work together with time-index modeling for data-driven glucose prediction. Besides, Chapter 4 will also present the effectiveness of applying the traditional model-based transfer learning technique to the self-attention-based model via pre-training & fine-tuning.

2.3 Engineering

This section reviews the existing studies related to the engineering of data-driven glucose prediction. In this context, engineering is a more diverse topic concerning ML developments other than the layer definition which is the focus of the modeling topic in Section 2.1.1. Thus, Section 2.3.1 first defines the coverage of the objective "engineering" in this thesis and gives examples in generic ML and DL, then Section 2.3.2 moves into the review specific to data-driven glucose prediction.

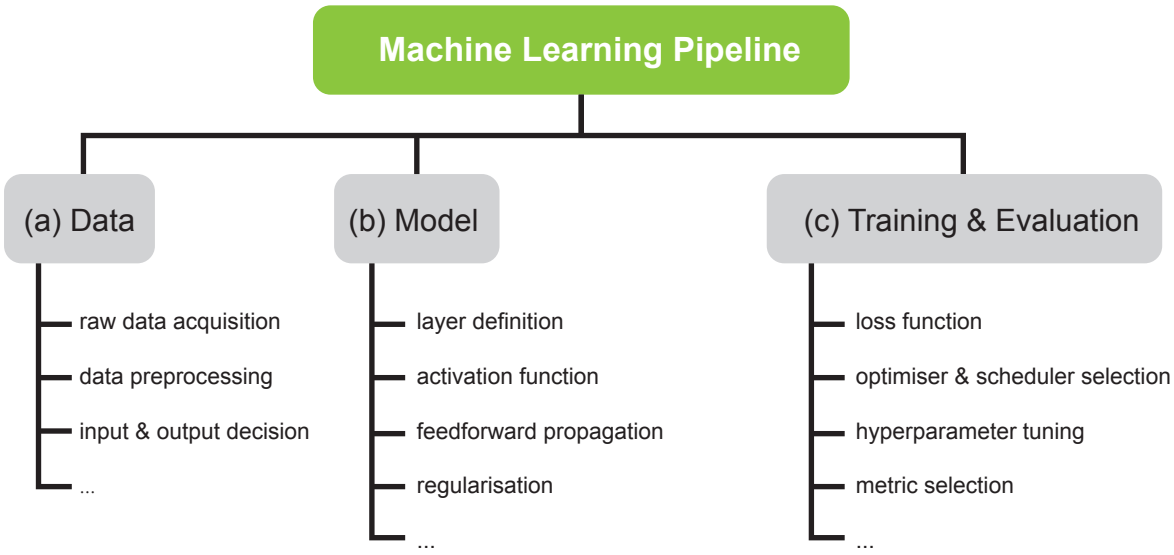


Figure 2.6: Components of a supervised learning pipeline. In this sketch, a machine learning pipeline is broken down into three segments, namely data, model, and training & evaluation. Among these segments, the data segment concerns *what* the ML system processes, while the model segment and the training & evaluation segment concern *how* the system processes it.

2.3.1 Engineering in ML

Section 2.1.1 has reviewed the recent development of ML and data-driven glucose prediction specifically in the modeling aspect, in which a taxonomy of RNN-based, CNN-based, Transformer-based, and time-index based models is employed. This taxonomy effectively covers the modeling part of a generic ML pipeline, dominantly the layer definition, as illustrated in Figure 2.6. This Section 2.3, on the other hand, reviews the developments related to other components in a ML pipeline, from data to training and evaluation, which is referred to as the engineering objective in this thesis. To make a further distinction for better understanding the scope, the data segment concerns the definition and manipulation of the input and output of the ML system, i.e., *what* the system processes and predicts, while the segments for the model and its training and evaluation concern *how* they are processed. Due to the diversity of this scope, developments to be reviewed in this section are more individual techniques, ML engineering craft, or even "tricks" as in Chen, Wilson et al. (2015); Jang et al. (2016), that can be employed in a ML pipeline ad hoc. To this end, this section does not aim at reviewing the large topic of engineering systematically, but rather gives concrete examples of engineering contributions in generic ML and DL in order to demonstrate the definition and coverage of the engineering objective that this thesis concerns.

In the data segment (Figure 2.6(a)), a common engineering technique is the feature scaling which is broadly used in almost all applications in ML, including my studies in Chapter 3-7. *Feature scaling* techniques like normalisation and standardisation are used to scale the features of the data to a specific range (e.g., 0 to 1) or to have zero mean and unit variance. Without feature scaling, different features that have different original distributions would become a bottleneck for the training of many ML algorithms. Features with larger values can dominate the learning process, causing the model to place undue importance on these features while neglecting smaller-valued features. Another engineering technique is the *data augmentation*, which specifically deals with limited volume of data. Data augmentation involves creating additional training data from existing data by applying various transformations. A representative application of data augmentation is the ImageNet classification task conducted by

Krizhevsky et al. (2012), in which horizontal flipping, random cropping, and RGB channel variation are applied on the original images in the dataset. This technique helps to increase the diversity of the training set, reduce overfitting, and improve the model generalisation to new data. *Data imputation* is another engineering technique that is widely used in time series forecasting. The natural difficulty in collecting time series data, due to the glitches of the data collection device or other disruptions, makes occasional missing entries a common issue in time series dataset. Data imputation methods are employed to fill in these missing values, ensuring the continuity and integrity of the dataset. Common techniques for data imputation include mean or median imputation, forward or backward filling, and more sophisticated methods like k-nearest neighbors imputation. In more recent Transformer-based time series forecasting methods, *input patching* becomes an important engineering practice in order to mitigate the computational load brought by the large Transformer model (Das et al., 2023). By using input patching, the time series is no longer processed in a token-by-token manner, but several adjacent tokens are treated as a patch. This idea is inspired by the image patching (Dosovitskiy et al., 2020) in computer vision, in which an image is treated as a sequence of small image patches. To summarise, data processing sets the foundation for the subsequent model design segment and training & evaluation segment, by ensuring that the input data is clean, consistent, and in a suitable format for analysis.

In the model segment (Figure 2.6(b)), engineering techniques are more in the form of a specific layer or operation in the neural network. The most representative engineering technique is probably the *residual connection* (He et al., 2016). In the feedforward process of a certain layer $f(\mathbf{x}) = \mathcal{H}(\mathbf{x})$ in a deep neural network, residual connection is simply adding a summation of the input $\mathbf{x}$ to the original output $\mathcal{H}(\mathbf{x})$, such that the mapping is recast into $f(\mathbf{x}) = \mathcal{H}(\mathbf{x}) + \mathbf{x}$. This operation does not add extra complexity due to the simplicity of the summation operation, and does not essentially change the learning capacity of this layer, as it only changes the learning objective from an underlying true mapping $\mathcal{F}(\mathbf{x})$ into $\mathcal{F}(\mathbf{x}) - \mathbf{x}$. However, it creates a shortcut for the gradients to smoothly flow back by bypassing the complex $\mathcal{H}(\mathbf{x})$ and treating this layer $f(\cdot)$ as an identity mapping. The residual connection is originally applied on image recognition, but soon demonstrates itself a generic engineering technique effective in any deep network. Moreover, *layer normalisation* (Ba et al., 2016) is another engineering technique that is specifically useful in RNNs and Transformers. In layer normalisation, the output of a certain layer in the neural network is normalised by subtracting the mean and dividing the standard deviation. Then two trainable parameters are applied to scale and shift back the distribution of the normalised output. By normalising the inputs of each layer in a neural network, layer normalisation helps reduce the internal covariate shift, which is the change in the distribution of network activations due to changes in network parameters during training. As a result, this operation leads to faster convergence and more stable training, and has been used as a default engineering technique in the Transformer (Vaswani et al., 2017). To summarise, it can be seen that these engineering techniques are modules that can be incorporated in the model design rather than complete models, thus revealing the distinction between the engineering objective to the modeling objective.

In the training & evaluation segment (Figure 2.6(c)), many engineering techniques focus on more effective training and improving model generalisation. For example, *batch normalisation* (Ioffe and Szegedy, 2015) is an effective approach to improve the training speed and stability of deep neural networks. By normalising the inputs of each mini-batch, batch normalisation addresses the issue of internal covariate shift, which is the change in the distribution of network activations due to updates in the parameters during training. Note the distinction between layer normalisation and batch normalisation, where layer normalisation is a module used solely in the model, while batch normalisation is only used in the training phase targeting a specific mini-batch. This normalisation allows the use of higher learning rates and

reduces the sensitivity to network initialisation. Another engineering technique with similar purpose is *dropout* (Srivastava et al., 2014). In each training iteration using a mini-batch of data, this technique randomly sets the outputs of a small portion of neurons in the neural network to zero, thus "dropping out" those neurons and their corresponding computations of the gradient from training. In the evaluation phase, the dropout is turned off and the full set of neurons are used in model inference. Intuitively, the operation of muting some neurons in the training could improve the robustness of the network by ensuring that it does not rely heavily on any single neuron. Therefore, this process forces the network to learn redundant representations, thus reducing the chance of overfitting. More engineering techniques enhancing the training & evaluation include *early stopping* (Prechelt, 2002), which halts training when performance on a validation set ceases to improve, *learning rate scheduling* (Smith, 2017), which adjusts the learning rate dynamically during training, and *gradient clipping* (Zhang, He et al., 2019), which addresses the issue of exploding gradients by capping gradients at a maximum threshold. These methods collectively contribute to more effective training regimes and robust model generalisation, addressing various challenges inherent in deep neural network training.

This section has reviewed the most successful engineering techniques from a generic ML perspective. These reviewed techniques reflect the scope of the engineering objective in this thesis, i.e., techniques that less relate to each other, yet collectively contribute to building an effective ML pipeline.

2.3.2 Engineering in Data-driven Glucose Prediction

Section 2.3.1 has surveyed a series of engineering techniques and has revealed the scope of the engineering objective in this thesis. This section reviews the same topic, while zooming into the field of data-driven glucose prediction.

In the data segment, data imputation is widely adopted in existing studies of data-driven glucose prediction as a data preprocessing technique. The missing entry is a very common issue in collecting CGM time series data, caused by device glitches such as the occasional loss of Bluetooth signal (Kim et al., 2020). Different data imputation techniques like linear interpolation (Zhu, Li, Herrero, Chen et al., 2018; Zhu, Kuang et al., 2024), spline interpolation (Pérez-Gandía et al., 2010) have been employed in existing studies to maximise the usability of the CGM data. Jeon et al. (2020) evaluated a series of traditional imputation methods in this context, and concludes that the ensemble of five models achieves their best predictive performance, where the models are trained respectively on CGM data imputed by carrying forward the last observed value, linear interpolation, spline interpolation, Stineman interpolation, and Kalman smoothing. Acuna et al. (2023) evaluate different imputation methods for CGM data including hourly mean, linear interpolation, cubic interpolation, spline interpolation, Kalman smoothing, and smoothing splines. Their comparative experiments show that linear interpolation achieves the best imputation method using the Transformer model. These findings emphasise the importance of selecting appropriate imputation techniques to enhance the accuracy and reliability of CGM data for subsequent predictive modeling tasks.

The input & output decision is another engineering concern specifically for glucose prediction in the postprandial scenario. Existing research has mainly approached postprandial glucose prediction from two directions, namely prediction at meal time (Oviedo, Contreras et al., 2019; Pustozerov et al., 2020; Vehí et al., 2020; Adelberger et al., 2021) and prediction within a fixed horizon (Seo, Lee et al., 2019; Karim et al., 2020). Meal-time prediction considers a pre-defined meal effective time range such as 4 hours after the meal, and all predictions regarding the time range are made at once at the meal time. This method reflects the application that

the patient could re-plan the meal advised by the prediction result, e.g., adjust the planned food quantity or insulin treatment. Although the envisaged application is of great value, a natural disadvantage of this method is that it cannot take factors after the meal intake into consideration. In contrast, the fixed prediction horizon approach considers a shorter prediction horizon such as 30 or 60 minutes, and the 30 or 60 minutes horizon is rolling, such that the whole of the 4 hour meal event is covered. This approach corresponds to the application that once a meal event is announced by the patient, the algorithm constantly runs to predict glucose for a short future, such that later events that reflect on glucose profile can be accommodated. Although having a shorter prediction horizon than the meal-time prediction approach, the application of fixed horizon prediction is more flexible. Regarding this engineering concern, a study of mine including a refined input & output decision for postprandial glucose prediction will be presented in Chapter 6. Another divergence in the input & output decision is the prediction target being single-point (Dong et al., 2019; Li, Liu et al., 2019; Mirshekarian et al., 2019) or the full excursion (Calm, Garcia-Jaramillo et al., 2007; Adelberger et al., 2021), which will be discussed further in Chapter 7.

In the model segment, one engineering technique lies in the way of making predictions after encoding the input glucose sequence into a deep representation. To predict the glucose at a prediction horizon which is multiple time steps ahead, the direct prediction technique (Fox et al., 2018; Kim et al., 2020; Rabby et al., 2021; Koca et al., 2023) takes the representation to a direct mapping (usually a linear mapping) to compute the predicted glucose. This method maximally simplifies the subsequent prediction process, making it computationally efficient and easier to implement. By directly mapping the encoded representation to the predicted glucose values, the model reduces the complexity and the number of parameters required, which can lead to faster training and inference times. In contrast, another approach is the rolling prediction technique (De Bois, Yacoubi et al., 2019; Zaidi et al., 2021). Different to direct prediction, the encoded representation is is used to predict glucose at the next step, rather than the final target. The prediction is then appended into the input glucose sequence to enable the prediction of one step further. By following the temporal pattern, this recursive rolling prediction can better capture temporal dependencies in the glucose data. Another engineering technique in the model segment is the use of exogenous inputs other than the CGM sequence, e.g., insulin treatments and food intake (Canete et al., 2012; Fong et al., 2013; Efendic et al., 2014; Zhu, Li, Herrero, Chen et al., 2018; Li, Daniels et al., 2019). Despite the fact that many studies use multi-model inputs, there are studies suggesting that including such information might be redundant (Lu et al., 2010; Oviedo, Vehí et al., 2017), and some recent studies use uni-modal input of CGM only (Fox et al., 2018; Martinsson, Schliep, Eliasson and Mogren, 2020). In conclusion, existing research presents divergent views on these engineering techniques, highlighting the need for further investigation to determine the most effective strategies for glucose prediction modeling. Detailed discussion regarding this inconsistency will be presented in Chapter 7.

The above review highlights the existing engineering techniques in data-driven glucose prediction. In the field of data-driven glucose prediction, the dominant research focus remains on the modeling as reviewed in Section 2.1.2. Consequently, few consensuses have been reached regarding the best engineering practices, resulting in considerable variation among existing studies in terms of engineering details like data processing and model feedforward. This will be specifically discussed in Chapter 6 and 7.

2.4 Summary

In this chapter, we have reviewed the current state of the art in data-driven glucose prediction, focusing on three key areas: modeling, adaptation, and engineering. The exploration of these aspects has provided insights into the existing techniques and the challenges that remain in this evolving field.

Modeling Various approaches, including RNNs, CNNs, and Transformers, have been employed in data-driven glucose prediction. While RNNs and their variants such as LSTMs and GRUs have been widely used due to their ability to handle sequential data, they often struggle with issues like vanishing and exploding gradients and difficulty in training. CNNs, adapted for time-series data, provide a more computationally efficient alternative but the use of local kernels may lack the learning capacity. The recent introduction of Transformer-based models offers a promising direction due to their ability to learn complex dependencies and relationships in data. However, these models require large amounts of data for effective training, which can be a limitation in glucose prediction. A newer approach using time-index models, such as the MOTIM, has recently been proposed and evaluated in generic time series forecasting, but have not yet been applied to data-driven glucose prediction. My research on the initial attempt of using the Transformer model and its self-attention mechanism is to be presented in Chapter 4, and the research of applying the MOTIM will be presented in Chapter 5.

Adaptation The adaptation of models to individual patient data remains a critical challenge. Techniques such as transfer learning and meta-learning have been proposed to address the issue of limited personalised data. While these methods show promise, there is still a need for more refined approaches that can efficiently adapt generic models to specific patient needs with minimal data. This aspect is particularly important given the variability in glucose dynamics across different individuals. The challenges and advancements in adaptation are further explored in Chapters 4 and 5, where novel adaptation techniques and their integration with advanced modeling methods for data-driven glucose prediction are evaluated and discussed.

Engineering From an engineering perspective, the practical implementation of these models requires a balance between accuracy and computational efficiency. The trade-off between model complexity and usability in real-time applications is a key consideration. Additionally, comparing the different engineering settings observed in the literature is another open research direction that could benefit the field by identifying the most effective strategies for implementation. Efficient and effective data preprocessing, training, and testing strategies are also essential to enhance the performance and usability of glucose prediction systems in real-world settings. Targeting the above issues, Chapter 5 discusses how the MOTIM method improve the model efficiency by the use of its time-index modeling; Chapter 6 addresses the refined training and evaluation strategy specifically designed for data-driven glucose prediction in the postprandial scenario; Chapter 7 evaluates and discusses the prediction schemes and prediction targets that existing studies on data-driven glucose prediction have adopted, in order to identify the most effective approaches for optimising both predictive accuracy and computational efficiency in real-world applications.

This chapter has reviewed the related works from the perspectives of modeling, adaptation, and engineering. Next, Chapter 3 to 7 will present my individual studies, with each chapter corresponding to a specific study.

Chapter 3

Blood Biomarkers Importance Assessment

This chapter has been published as

R. Cui, E. Daskalaki, M. Z. Hossain, A. Lenskiy, C. J. Nolan and H. Suominen (2021). 'A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning'. In: *2021 15th International Congress in Nursing Informatics (NI)*. IOS Press, pp. 36–38. DOI: 10.3233/SHTI210657

Author contributions: R.C., E.D., and H.S. conceptualised the study; R.C. designed and performed research; R.C. wrote the manuscript and revised it, based on critical comments by E.D., M.Z.H, A.L., C.J.N., and H.S.

This chapter presents a study on blood biomarkers importance assessment using ML, which serves as a preliminary exploration on the application of ML in analysing diabetes. The 2009 China Health and Nutrition Survey (CHNS) (Popkin et al., 2010) provides blood samples of 9,549 individuals, containing data of 30 blood biomarkers including Fasting Plasma Glucose (FPG) and Hemoglobin A1C (HbA1c), together with their age and gender information. These data were used to conduct the diagnosis of diabetes, which was formulated as a classification task using ML. By alternatively using the FPG and HbA1c as the ground truth in labeling the presence of diabetes and rest of the remaining information as the set of input features, the employed XGBoost (Chen and Guestrin, 2016) algorithm achieved over 80% F1 scores. Next, a permutation feature importance test (Fisher et al., 2019) was performed to study the contribution of each feature towards the classification. It was concluded that the rest of the blood biomarkers were no more important than age and blood insulin, in their correlation with the presence of diabetes.

3.1 Introduction

Diabetes diagnosis and discovery of important biomarkers is an active research field, which has been investigated using ML techniques (Elsherbini et al., 2022; Ortiz-Martínez et al., 2022). Various ML algorithms have been investigated in diagnosing diabetes such as Adaboost, logistic regression, decision trees, random forests, and neural networks (Vijayan and Anjali, 2015; Zou et al., 2018; Lai, Huang et al., 2019). This approach of using ML to diagnose diabetes has made significant contributions to the field, improving diagnostic accuracy and enhancing our understanding of diabetes. For example, 8-hydroxy-2-deoxyguanosine

Table 3.1: Constructed datasets statistics.

Labelling scheme	FPG $\geq$ 7.0 mmol/l	HbA1c $\geq$ 6.5%
Total No. of positive examples	550	642
Total No. of negative examples	8,725	8,633
No. of constructed negative datasets	15	13
No. of features	31 (FPG excluded)	31 (HbA1c excluded)
Class balance of constructed datasets	550 positive, 550 negative	642 positive, 642 negative

[a] Abbreviations: FPG, Fasting Plasma Glucose; HbA1c, Hemoglobin A1C.

has been identified as a co-marker of HbA1c by the use of decision trees (Jelinek et al., 2016), and the relationship between 5'AMP-activated protein kinase and diabetes has been identified using random forests (Huang et al., 2013). Following this approach, this study investigated the correlation of a list of blood biomarkers to the presence of diabetes[1]. In terms of determining the presence of diabetes, according to the World Health Organisation (WHO), diabetes can be diagnosed by FPG $\geq$ 7.0 mmol/l (World Health Organization and International Diabetes Federation, 2006), while HbA1c $\geq$ 6.5% is considered as an alternative by the International Expert Committee (The International Expert Committee, 2009), and also recommended by WHO (World Health Organization, 2011). Although FPG and HbA1c exhibit a proven strong correlation (Khan et al., 2007), the two diagnostic criteria are not always equivalent leading often to contradicting outcomes (Sacks, 2011).

The objective of this study was to explore the differences between the two aforementioned criteria and investigate the correlations of the selected blood biomarkers with the diagnosis of diabetes. Various traditional ML algorithms were used for diagnosing diabetes using the measurements of the blood biomarkers in the CHNS dataset as the input, where the aforementioned diagnostic criteria were alternatively applied as the definition of diabetes. In these algorithms, the XGBoost algorithm achieved an over 80% F1 score. To investigate the correlation and contribution of the included biomarkers to the diagnosis, a permutation features importance test was subsequently conducted. The main outcome was that a high FPG or HbA1c was the most significant biomarker for the presence of diabetes, while insulin was informative when diabetes was diagnosed by FPG, but insignificant when diabetes was diagnosed by HbA1c.

3.2 Methodology and Experiments

This section introduces the approach of preprocessing the dataset in Section 3.2.1, while the ML configuration using the data is presented in Section 3.2.2.

3.2.1 Data Preprocessing

This study was based on data collected in the framework of CHNS (Popkin et al., 2010), which contains the blood samples from over 9,549 individuals. After discarding entries with missing values, the final dataset had 9,275 subjects, each equipped with 32 features[2]. The

[1] Details can be found in Table A.1.

[2] The 32 features were: age, gender, UREA, UA, APO_A, LP_A, HS_CRP, CRE, HDL_C, LDL_C, APO_B, MG, FET, INS, HGB, WBC, RBC, PLT, FPG, Y48_2, Y48_3, Y48_4, Y48_5, Y48_6, HbA1c, TP, ALB, TG, TC, ALT, TRF, TRF_R.

subjects living with diabetes were identified by the two diagnostic criteria: FPG $\geq$ 7.0 mmol/l and HbA1c $\geq$ 6.5%, and two groups of 550 and 642 subjects with diabetes were respectively labelled. It was noteworthy that only 351 subjects met both criteria. To aid the learning, 20 non-Gaussian features were log-transformed (Feng et al., 2014) to fulfil the Gaussian distribution assumption for algorithms like logistic regression[3]. Moreover, all features were normalised to the range [0, 1] by subtracting the minimum value and dividing by the gap between the maximum and the minimum value. For each labelling case, the biomarker used as the label, i.e., FPG or HbA1c, was then discarded from the input features set.

Since the data were largely unbalanced where the number of subjects without diabetes was dominantly larger than those with diabetes, the following procedure was employed to make the maximal use of the data. As shown in Table 3.1, taking labeling criterion $FPG \geq 7.0 mmol/l$ as an example, 550 of the subjects met the criterion and were labeled as positive. By sampling the same number of subjects from the negative subjects, 15 sets of 550 different subjects were sampled as 15 constructed datasets. A training round was based on taking the positive set and one of the 15 constructed negative datasets as the data. In the 1,210 subjects, a stratified 10-fold experiment was then performed, meaning 1/10 of these subjects were treated as the test set. For the two labeling criteria, 15 and 13 constructed negative datasets were able to be defined. The reported scores in this study were thereby the mean of the 150 and 130 training rounds.

3.2.2 ML Formulation

The problem of diabetes diagnosis was formulated as a classification task using the blood biomarkers. The following four ML algorithms were explored: Logistic Regression (LR), Support Vector Machine (SVM), XGBoost (XGB), and Multilayer Perceptron (MLP). F1 score was employed as the metric for performance evaluation. In addition, considering that the importance of features in the classification task would imply a clinical significance of the corresponding biomarkers in diabetes diagnosis, a permutation feature importance evaluation (Fisher et al., 2019) was conducted. This technique quantified the importance of a certain feature as the amount of performance decrease after ruining the semantic of that feature, i.e., randomising the data of that feature in the dataset while preserving its mean and variance. To enable more detailed comparison, two sets of experiments were defined, which were differentiated by the selection of input features. More specifically, the *experiment 1* (E1) included all biomarkers as input features, i.e., containing either FPG or HbA1c, while the *experiment 2* (E2) included only the biomarkers which were common between the two labelling cases, i.e., no FPG or HbA1c in the feature set. In both E1 and E2, age and gender were added into the feature set to include demographic information as input. For further enhancing the reliability of the results, all experiments on each constructed dataset were performed based on 10-fold cross validation.

3.3 Results

This section presents the results of the ML based diabetes diagnosis experiments and the feature importance evaluation. The results for the diagnosis are presented in Table 3.2. Among the four algorithms, best results were achieved by XGBoost. For the case of E1, the best classification performance was in the range of 86% – 89%. The performance experienced a drop in E2, yet still reached approximately 80% when both FPG and HbA1c were excluded.

[3]Non-Gaussian means the distribution shown in the dataset is not Gaussian-like. The 20 non-Gaussian features were: UREA, UA, APO_A, LP_A, CRE, HDL_C, MG, FET, INS, WBC, RBC, FPG, Y48_2, Y48_5, Y48_6, HbA1c, TG, TC, ALT and TRF_R.

Table 3.2: Results of the two classification experiments with four ML algorithms for each experiment. All the presented values are over all constructed datasets and validation folds, with the best scores highlighted in bold.

labelling cases		FPG $\geq$ 7.0 mmol/l				HbA1c $\geq$ 6.5%			
model		LR	SVM	XGB	MLP	LR	SVM	XGB	MLP
E1	Accuracy	0.8615	0.8826	**0.8908**	0.8805	0.8518	0.8612	**0.8649**	0.8546
	F1 score	0.8584	0.8757	**0.8888**	0.8754	0.8444	0.8549	**0.8629**	0.8521
E2	Accuracy	0.8081	**0.8312**	0.8235	0.8209	0.7958	**0.7980**	0.7939	0.7898
	F1 score	0.8050	**0.8243**	0.8215	0.8158	0.7884	0.7917	**0.7919**	0.7873

[a] Abbreviations: FPG, Fasting Plasma Glucose; HbA1c, Hemoglobin A1C; LR, Logistic Regression; SVM, Support Vector Machine; XGB, XGBoost; MLP, Multi-layer Perceptron.

This was as expected since the exclusion of FGP and HbA1c was a significant loss of information. For the feature importance evaluation (Table 3.3), all ML algorithms were consistent in terms of the most significant features identified. E1 indicated that FPG (HbA1c) was the most significant biomarker when using HbA1c (FPG) as the diagnostic criterion. The second most important biomarker for the labelling case of FPG $\geq$ 7.0 mmol/l was insulin, which conversely presented very low significance for the labelling case of HbA1c $\geq$ 6.5%. Moreover, from E2, where both FPG and HbA1c were excluded, age demonstrated its significance in both labelling cases, but it was not as insightful as FPG, HbA1c, and insulin, due to a relatively lower importance value. The impact of age factor in determining the presence of diabetes coincided with the clinical sense that diabetes is more often seen in older ages. Note that the presence of diabetes in this study is determined by the aforementioned diagnostic metrics, hence no distinction of type 1 diabetes and type 2 diabetes can be made.

3.4 Discussion

This study investigated diabetes diagnosis from blood biomarkers by applying ML classification algorithms to blood samples collected from a generic Chinese population. Using two different diagnostic criteria, FPG $\geq$ 7.0 mmol/l and HbA1c $\geq$ 6.5%, the best F1 scores were 89% and 86.5%, respectively. The FPG and HbA1c were mutually the most significant diabetes biomarker when the other was used as the ground truth for diagnosis. This agreed with the evidence of both FPG and HbA1c being crucial markers for diabetes diagnosis. Although insulin was an important marker when diabetes was diagnosed by FPG $\geq$ 7.0 mmol/l, it presented no significant correlation with HbA1c in this study. An explanation is that HbA1c, as an indicator of average glycemia during the past three months, presented lower correlation with insulin than FPG. However, note that for patients receiving external insulin therapy, their insulin levels may not solely reflect their body's endogenous insulin production but also the externally administered insulin. This could potentially affect the correlation between insulin levels and the presence of diabetes. Consequently, while insulin remains a crucial factor in diabetes management, its direct correlation with diabetes status may be confounded by external insulin administration. Besides, age appeared in high importance for the algorithms, which confirmed its strong position as a risk factor.

The main limitation of this study was that the CHNS study was not originally focused on diabetes. Despite its large cohort, not many individuals were with diabetes. This rendered

Table 3.3: Sorted average permutation features importance [%] in the two experiments where each value is the average importance over all constructed datasets and validation folds. As an example, a feature importance value of 5.72% indicates that permuting this feature resulted in decrease of the performance by 5.72%. The features whose absolute value are $\geq 1\%$ are highlighted in bold.

		LR		SVM		XGB		MLP	
E1	FPG	**HbA1c**	**5.72**	**HbA1c**	**5.43**	**HbA1c**	**6.96**	**HbA1c**	**5.76**
		INS	**2.64**	**INS**	**2.67**	**INS**	**3.32**	**INS**	**2.70**
		Y48_4	0.36	MG	0.33	age	0.40	TP	0.34
		LP_A	0.32	Y48_4	0.30	Y48_3	0.28	ALB	0.28
		...	...	...	...	...	...	...	...
		HGB	-0.15	Y48_5	-0.07	Y48_2	-0.22	HDL_C	-0.13
	HbA1c	**FPG**	**6.05**	**FPG**	**6.35**	**FPG**	**7.04**	**FPG**	**6.43**
		age	0.73	age	0.43	ALT	0.31	age	0.50
		APO_A	0.27	APO_B	0.42	TP	0.29	Y48_3	0.36
		APO_B	0.22	TP	0.27	APO_A	0.29	TP	0.27
		...	...	...	...	...	...	...	...
		Y48_4	-0.16	WBC	-0.06	PLT	-0.09	LP_A	-0.16
E2	FPG	**INS**	**2.20**	**INS**	**2.83**	**INS**	**2.85**	**INS**	**2.26**
		age	0.82	**age**	**1.10**	**age**	**1.42**	age	0.72
		LP_A	0.40	ALB	0.83	ALB	0.57	UA	0.38
		Y48_4	0.37	UA	0.80	UA	0.52	ALB	0.32
		...	...	...	...	...	...	...	...
		TG	-0.24	TRF_R	0.00	LDL_C	-0.07	TG	-0.44
	HbA1c	**age**	**2.04**	**age**	**1.39**	**age**	**1.40**	**age**	**1.43**
		INS	0.63	APO_B	0.48	MG	0.59	Y48_3	0.40
		Y48_3	0.58	RBC	0.44	APO_B	0.48	APO_B	0.39
		APO_B	0.48	INS	0.43	HS_CRP	0.41	INS	0.37
		...	...	...	...	...	...	...	...
		HDL_C	-0.02	TRF_R	-0.12	TRF	-0.35	UREA	-0.34

[a] Abbreviations: FPG, Fasting Plasma Glucose; HbA1c, Hemoglobin A1C; LR, Logistic Regression; SVM, Support Vector Machine; XGB, XGBoost; MLP, Multi-layer Perceptron; more to be found in Appendix A.

the classification problem unbalanced and, with this current approach, could not take full advantage of the large study population. Another limitation was the fact that stratified

sampling could change the original distribution. As a result, the trained ML algorithm using the constructed datasets can only be applied for theoretical analysis of the feature importance, but not real-world diagnostic application. Hence, future work could be applying bespoke ML methods for class unbalance.

Continuing from this initial exploration of using ML to analyse blood glucose as a biomarker, I moved to the main topic of this thesis, i.e., data-driven glucose prediction. My four studies in this field are to be presented in the rest of this thesis, namely Chapter 4, 5, 6 and 7.

Chapter 4

Glucose Prediction via Recurrent Self-attention Network

This chapter has been published as

R. Cui, C. Hettiarachchi, C. J. Nolan, E. Daskalaki and H. Suominen (2021). 'Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network'. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. DOI: 10.1109/CBMS52027.2021.00064

The corresponding work is open-sourced at https://github.com/r-cui/GluPred.

Author contributions: R.C., C.H., C.J.N, E.D., and H.S. conceptualised the study; R.C. and C.H. designed and performed research; R.C. wrote the manuscript and revised it, based on critical comments by C.J.N., E.D., and H.S.

This study targeted the objective "modeling" and "adaptation" for data-driven glucose prediction, as introduced in Section 1.2.1 and 1.2.2. The focus was on 1) how can the existing models be improved in terms of predictive performance, and 2) how the improved model performs when combining with transfer learning technique for the adaptation purpose. The presented study was based on historical-value model, which learned a mapping from recent glucose sequence to the future value. At the time of this study, a remarkable advancement of deep learning was the Transformer (Vaswani et al., 2017), which was shown effective in other sequence modeling tasks such as natural language processing. Inspired by this, the presented study investigated the research question of whether it could be used to improve existing historical-value based glucose prediction models. I employed the self-attention module of Transformer and proposed a novel model, Recurrent Self-attention Network (RSAN), for data-driven glucose prediction. The experiments using the OhioT1DM dataset achieved state-of-the-art performance with average RMSE results of 17.82 mg/dL for predicting 30 minutes ahead and 28.54 mg/dL for predicting 60 minutes ahead. These results indicated the effectiveness of the proposed modeling strategy for the task of short-term glucose prediction.

4.1 Introduction

To date, existing methods for data-driven glucose prediction are based on historical-value modeling, which learns a mapping from the recent past glucose sequence to the targeted future values. Previous studies have employed RNNs and CNNs as the backbone (Allam et al., 2011; Sun et al., 2018; Li, Daniels et al., 2019; Kim et al., 2020; Martinsson, Schliep,

Eliasson and Mogren, 2020; Zhu, Li, Chen et al., 2020). As reviewed in Section 2.1, the two main types of modeling differ in the approach to encode the glucose sequence into deep features, where RNNs follow an autoregressive manner, while TCNs use cascaded moving kernel. There exists limitations for both structures — RNNs suffer from gradient vanishing and exploding problems (Pascanu et al., 2013) which brings difficulty in training; TCNs are limited in learning the distanced relations in the input sequence as the lowest-level kernel only captures the local pattern within the receptive field of length d. Therefore, the encoding process, reflected by the model structure design, holds the potential to be further improved. At the time of this study, the Transformer model (Vaswani et al., 2017) had demonstrated promising performance in a variety of fields as reviewed in Section 2.1.1, including natural language processing (Devlin et al., 2018; Liu, Ott et al., 2019; Ott et al., 2019; Yang, Dai et al., 2019), computer vision (Dosovitskiy et al., 2020; Fang et al., 2021; Liu, Lin et al., 2021), and time series forecasting (Liu, Yu et al., 2021; Wu et al., 2021; Zhou et al., 2021). In light of the success of transformers in these fields, it is reasonable to hypothesise that it also holds potential in the glucose prediction task.

The objective of this study was to investigate if the Transformer could indeed be applied in the data-driven glucose prediction task and improve the performance. To achieve this, I employed the core module of Transformer, i.e., the self-attention module, and proposed the RSAN as a novel model in this field. Moreover, in training the RSAN model, transfer learning was employed to make use of data from other patients to aid the learning of a personalised model. The RSAN was evaluated using the OhioT1DM dataset (Marling and Bunescu, 2020) which is a standard benchmark in the field, and achieved state-of-the-art performance with average RMSE results of 17.82 mg/dL for 30 minutes and 28.54 mg/dL for 60 minutes.

4.2 Methodology

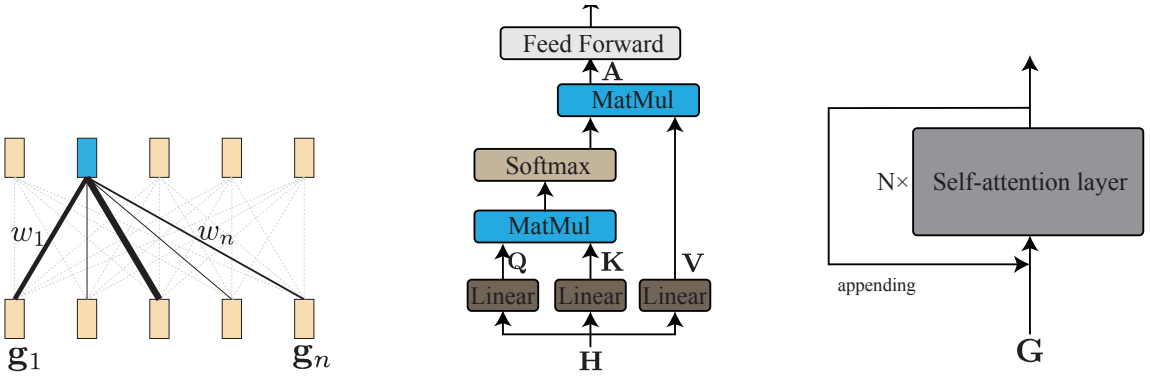
This section introduces the methodology in detail. The problem is first formally defined in Section 4.2.1. The RSAN model stacked several self-attention layers, and the concept of a single self-attention layer is introduced in Section 4.2.2. The ensemble of the RSAN model is presented in Section 4.2.3. Section 4.2.4 presents the way I used the transfer learning technique to aid the training and mitigate the learning difficulty due to data amount limitation.

4.2.1 Problem Definition

In this study, the problem was defined as a single-point prediction problem following many existing studies in this field (Oviedo, Vehí et al., 2017), i.e., predicting a single-point future value at a certain horizon. Given an input sequence $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N]$ where $\mathbf{g}_i \in \mathbb{R}^F$ with F being the number of features of the sequence, the target was to predict the future value $\mathbf{g}_{N+T}$ where T was referred to as the prediction horizon. When applying this definition specifically to glucose prediction, the sequence features can include glucose and potentially other covariants, such as the insulin and carbohydrate intake information, and the prediction target was only one of the F dimensions, i.e., the glucose dimension. To this end, the target was to obtain a learned mapping

$$f([\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N]) = g_{N+T}$$

where the unbold g_{N+T} denoted the glucose scalar without exogenous dimensions. To tackle this, the RSAN model was proposed and applied to model the mapping $f(\cdot)$, which is to be introduced next.



(a) Illustration of the attention. (b) Self-attention layer. (c) Recurrent self-attention network.

Figure 4.1: (a) A high-level illustration of a learned self-attention. The attention was essentially a $n \times n$ weights matrix in which each token in the upper level was assigned with a weight vector of length n that summed to 1.0. (b) Illustration of a single layer of the self-attention based encoder of RSAN, in which the softmax layer computed the attention matrix. Figure is adapted from Vaswani et al. (2017). (c) Illustration of the recurrency in RSAN which stacked multiple encoder layers. In the autoregressive prediction, a next glucose value was predicted at a forward pass of the model, then the value was appended to the input sequence to enable the next prediction step. Abbreviations: RSAN, Recurrent Self-Attention Network.

4.2.2 Self-Attention Layer

The RSAN model employed multiple self-attention layers (Vaswani et al., 2017) as illustrated in Figure 4.1(b). To formally define the concepts, an i-th self-attention layer took as input a matrix $\mathbf{H}^{(i)} \in \mathbb{R}^{N \times D}$ where N was the length of the input sequence and D was the hidden dimension. The input $\mathbf{H}^{(i)}$ was then linearly transformed to get three variables, i.e., the query $\mathbf{Q}^{(i)} \in \mathbb{R}^{N \times D}$, the key $\mathbf{K}^{(i)} \in \mathbb{R}^{N \times D}$, and the value $\mathbf{V}^{(i)} \in \mathbb{R}^{N \times D}$ by

$$\begin{aligned}\mathbf{Q}^{(i)} &= \mathbf{H}^{(i)}\mathbf{W}_q^{(i)} \\ \mathbf{K}^{(i)} &= \mathbf{H}^{(i)}\mathbf{W}_k^{(i)} \\ \mathbf{V}^{(i)} &= \mathbf{H}^{(i)}\mathbf{W}_v^{(i)}\end{aligned}$$

where each of $\mathbf{W}_q^{(i)}, \mathbf{W}_k^{(i)}$, and $\mathbf{W}_v^{(i)}$ was in $\mathbb{R}^{D \times D}$. The self-attention $\mathbf{A}^{(i)} \in \mathbb{R}^{N \times D}$ was then defined as

$$\mathbf{A}^{(i)} = \text{softmax}(\frac{\mathbf{Q}^{(i)}\mathbf{K}^{(i)\top}}{\sqrt{D}})\mathbf{V}^{(i)}. \tag{4.1}$$

The self-attention layer output a nonlinear transformation $\sigma^{(i)}(\cdot)$ of the self-attention $\mathbf{A}^{(i)}$. Here, a fully-connected feed-forward network was applied on the hidden dimension D with a ReLU activation (Nair and Hinton, 2010) as the source of nonlinearity. The final output of a self-attention layer was consequently defined as

$$\begin{aligned}\mathbf{H}^{(i+1)} &= \sigma^{(i)}(\mathbf{A}^{(i)}) \\ &= \text{ReLU}(\mathbf{A}^{(i)}\mathbf{W}_1^{(i)} + \mathbf{b}_1^{(i)})\mathbf{W}_2^{(i)} + \mathbf{b}_2^{(i)}\end{aligned}$$

where the output $\mathbf{H}^{(i+1)}$ remained in $\mathbb{R}^{N \times D}$.

Taking a closer look at this feedforward computation, it can be spotted that the core mechanism of the self-attention layer was the n^2 paired learnable attention, i.e., the softmax$(\frac{\mathbf{Q}^{(i)}\mathbf{K}^{(i)^\top}}{\sqrt{D}})$ term in Equation 4.1. This term offered the output $\mathbf{H}^{(i+1)}$ a full access of each token in the input with a learnable weighting, as illustrated in Figure 4.1(a), which was the key difference to the one-way encoding in RNNs or the small local pattern based encoding in TCNs.

4.2.3 Recurrent Self-Attention Network

To increase the capacity of the model, multiple self-attention layers were stacked to form the full encoder, as illustrated in Figure 4.1(c). Besides the encoder, the network made use of two trainable linear transformations to bridge data to different dimensions. Specifically, one linear operation was applied to map the input from low-dimensional feature space $\mathbb{R}^F$ to high-dimensional hidden space $\mathbb{R}^D$ before data entering the encoder; the other linear operation contrarily mapped the output of the encoder back to the prediction space $\mathbb{R}^F$ at the end of the feedforward computation. Apart from this, given the fact that the aforementioned transformations were all with respect to the hidden dimension, the sole encoder was not able to distinguish different positions in the sequence. To tackle this, sinusoidal positional encoding (Vaswani et al., 2017) was applied before the first layer of the encoder. Overall, the full RSAN network took as input $\mathbf{G} \in \mathbb{R}^{N\times F}$ and output $\hat{\mathbf{G}} \in \mathbb{R}^{N\times F}$, where $\hat{\mathbf{g}}_N \in \mathbb{R}^F$ was regarded as the prediction of $\mathbf{g}_{N+1}$.

The above process was able to predict a next-step future glucose. To predict T steps ahead, I augmented the above concepts to form the RSAN. The RSAN was a serial connection of T self-attention networks, where each latter self-attention network took as input the concatenation of the input and the one-step prediction of the former self-attention network. This recurrent process mimicked the RNN as in Figure 2.1. Following the standard practice of recurrent deep networks, I made the T networks sharing their weights. Since all the operations were differentiable with respect to the parameters, the full network could be trained using standard back-propagation.

4.2.4 Transfer Learning

Apart from the model structure, transfer learning was employed as part of the training pipeline to further increase the performance of the predictive system. As reviewed in Section 2.2.1, knowledge learned from other domains, i.e., the learned glucose pattern from a population other than the targeted patient, may hold the potential of aiding the learning of the targeted patient. To this end, I employed the parameter-transfer technique, where I split the training process into a pre-training stage using data from non-target subjects, and a fine-tuning stage using personal data of the target subject. With this technique, the model reserved the knowledge of generic glucose metabolism learned from non-target subjects, then made use of the knowledge to fast adapt to the target subject.

4.3 Experiments

This section presents the details for the experiments on the RSAN. The setups and designs of experiments evaluating the RSAN are presented in Section 4.3.1; Section 4.3.2 describes two ablation studies as further analysis of the proposed method; the results of the above experiments are discussed in Section 4.3.3.[1]

[1]Dataset details can be found in Appendix A. Implementation details can be found in Appendix B.

4.3.1 Experiment Setups

The proposed RSAN model was evaluated on the OhioT1DM dataset. In the experiments, the 2-hour CGM history together with glucose-related events including carbohydrate intake and basal and bolus insulin treatments in the same time range were taken as a 4-channel input of the RSAN model ($F = 4$). Two groups of evaluations using 30 minutes and 60 minutes as the prediction horizon were conducted. Regarding the channels other than the glucose channel, I performed experiments in the following two scenarios, inspired by Mirshekarian et al. (2019).

Glucose Prediction (Setting 1) In this setting, the model knew the ground truth of non-glucose data till the moment of each prediction step. Namely, the basal rate, and any bolus or carbohydrate intake events in the prediction horizon were directly provided to the model in both the training and testing stages. This "what-if" scenario answered questions of what the glucose level will be if any glucose-related event in the features set is planned to happen within the prediction horizon.

All Features Prediction (Setting 2) In contrast to setting 1, the model was not provided with any information in the prediction horizon for this scenario. This implied that the model needed to implicitly learn the life patterns of the subject, and predict the life events together with the glucose values to have a performance comparable to setting 1. A trained model under this scenario could serve as a fully diagnostic tool of what the glucose level will most likely be according to the two-hour historical information only.

The RMSE, as the most widely applied metric in data-driven glucose prediction, was used as a preliminary evaluation of the RSAN model[2]. Most specifically, the RMSE was calculated by

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{j=1}^{M}(\hat{g}_j - g_j)^2} \tag{4.2}$$

where $\hat{g}$ and g were the prediction and ground truth of glucose values measured in mg/dL, respectively, and M was the total number of predictions in the test set. To evaluate the statistical significance of the results, the one-sided paired T test with the level of significance set at $p = 0.05$ was employed. When comparing the proposed approach to the baselines, the subject-level RMSE was used as the two groups of samples in the T tests, and I tested if the results were statistically significantly lower than the best available baseline. When comparing the proposed approach to the ablation experiments to be presented next, the RMSE of all predictions on the test set was used as the two groups of samples in the T tests, and I tested if the results were statistically significantly lower than the counterpart results in the ablation experiments.

4.3.2 Ablation Studies

To further investigate the behaviour of RSAN, the following two ablation studies were then performed.

Ablation Study 1 In this ablation study, the aim was to evaluate the influence of using non-glucose features in the features set. To this end, I designed the experiment of reducing the features set to only glucose values, where the problem correspondingly reduced to a uni-modal time-series prediction problem, i.e., $F = 1$.

[2]A more comprehensive evaluation of the model using other metrics can be found in Chapter 5.

Ablation Study 2 I also evaluated the effects of using transfer learning, i.e., the effects of using data from other subjects to pre-train the model. The second ablation study was removing the effect of the pre-training phase. Practically, to maintain the same training iteration, the data of other subjects in the pre-training phase were replaced with the target subject's personal training data. This led to the full training process using the target subject's own data.

4.3.3 Results

This section presents and discusses the results of the experiments. Compared to the benchmarks for prediction horizons of 30 minutes and 60 minutes (Tables 4.1 and 4.2), the proposed approach achieved state-of-the-art performance with a statistically significant reduction of RMSE in both setting 1 and 2. Moreover, since glucose-related events (changing the basal rate, taking an extra insulin bolus and having food) took time to affect the glucose concentration, the difference of predictive performance between setting 1 and 2 was more noticeable in the 60 minutes prediction horizon than in the 30 minutes prediction horizon. Table 4.3 showed that the presence of insulin and carbohydrate intake information increased the performance of the proposed approach on most subjects. This observation demonstrated that RSAN was able to learn the effects of insulin and carbohydrates on glucose and make use of them during inference. Additionally, as shown in Table 4.4, results were better when transfer learning was applied, indicating that data from other people with T1DM helped the personalised learning. This can be attributed to this RSAN model requiring a large data set to overcome the performance limits during training. Reducing the model size and making it more efficient is one of the focus points of the next Chapter 5.

In the main experiments, the approach had better performance in setting 1 than setting 2 ($p < 0.05$) in the 60 minutes prediction horizon, while a similar result was not observed in the 30 minutes prediction horizon. This observation was consistent with Mirshekarian et al. (2019), indicating that knowing the glucose-related events was more helpful in longer-term glucose prediction. Therefore, it can be inferred that the prediction could be more accurate if the ongoing glucose-related factors are continuously provided to adjust the prediction. Besides, the better results of applying transfer learning than using only personal data in ablation study 2 suggested that (i) larger amounts of personal data can further improve the predictive performance, and (ii) data from other people could serve as data augmentation when personal data are insufficient for training a glucose predictive system.

4.4 Discussion

The contribution of this study is listed as follows:

- I proposed for the first time to apply the self-attention technique to data-driven glucose prediction, the novel recurrent self-attention network outperformed existing benchmarks using RNN and CNN based models.
- I demonstrated by the ablation experiment that a personalised RSAN model can benefit from training using other patients' data via a transfer learning technique.

In this study, I have introduced a novel approach RSAN for personalised short-term glucose prediction for people with T1DM based on a combined short history of glucose values and glucose-related events. The approach outperformed state-of-the-art performance compared to existing data-driven glucose prediction methods in the category of historical-value based models. One limitation of the proposed RSAN lay on the attention computation which was

of an n^2 level of complexity. This could limit its computational efficiency which is important when being deployed to real-life use. Therefore, the future work can concentrate on both the predictive performance and the efficiency of the model design. I hope that the concepts presented in this research, along with the openly accessible code resources, have the potential to enhance the perspective of current recurrent or convolutional modeling to a new stage.

The RSAN model introduced by this study belongs to the category of historical-value modeling. Continuing from this study, I moved beyond the scope of historical-value methods and turned to investigate the use of time-index modeling which learns a completely different mapping than historical-value models. This new methodology, which is novel to existing studies in glucose prediction, is to be presented in the next Chapter 5.

Table 4.1: Results of the 30 minutes prediction horizon. A bold value indicates that it is the lowest in its comparisons. A result marked by an asterisk * is statistically significantly ($p < 0.05$) lower than its best available baseline.

		RMSE (mg/dL)													
Subject		*540*	*544*	*552*	*567*	*584*	*596*	Avg (SD)	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
Martinsson, Schliep, Eliasson, Meijner et al. (2018)					-				19.5	19.0	16.5	24.2	19.2	22.0	20.07 (2.67)
Bertachi et al. (2018)					-				18.83	19.43	15.88	22.86	17.84	21.12	19.33 (2.45)
Zhu, Li, Chen et al. (2020)					-				18.6	**18.0**	15.3	22.7	17.6	21.1	18.88 (2.64)
Mirshekarian et al. (2019)	Set. 1				-				individual RMSEs not available						18.19
	Set. 2														18.74
Nemat et al. (2020)		20.98	17.66	16.30	20.52	21.62	17.45	19.09 (2.22)	-						
Zhu, Yao et al. (2020)		20.14	16.28	16.08	20.00	20.91	16.63	18.34 (2.23)	-						
Rubin-Falcone et al. (2020)		20.10	15.97	15.74	20.70	20.57	16.24	18.22 (2.08)	-						
My approach	Set. 1	**19.49**	15.75	15.68	**18.96**	**19.52**	15.94	**17.56** (1.59)*	**17.57**	18.34	14.84	21.69	16.02	20.08	18.09 (2.53)*
	Set. 2	20.24	**15.26**	**15.49**	19.72	20.23	**15.90**	17.81 (2.49)*	17.59	18.34	**14.64**	**21.31**	**15.93**	**20.03**	**17.97** (2.49)*

[a] Abbreviations: RMSE, Root Mean Square Error.
[b] Set. 1: exogenous inputs available; Set. 2: exogenous inputs unavailable.

Table 4.2: Results of 60 minutes prediction horizon. A bold value indicates that it is the lowest in its comparisons. A result marked by an asterisk * is statistically significantly ($p < 0.05$) lower than its best available baseline.

		RMSE (mg/dL)												
Subject		*540*	*544*	*552*	*567*	*584*	*596*	Avg (SD)	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
Martinsson, Schliep, Eliasson, Meijner et al. (2018)					-				34.4	29.9	28.6	37.3	33.1	36.0	33.22 (3.41)
Bertachi et al. (2018)					-				32.52	31.33	27.48	35.28	30.12	33.60	31.72 (2.74)
Mirshekarian et al. (2019)	Set. 1				-				individual RMSEs not available						29.12
	Set. 2														30.63
Nemat et al. (2020)		39.05	30.42	29.38	36.52	37.01	28.92	33.55 (4.46)	-						
Zhu, Yao et al. (2020)		38.54	27.64	29.03	35.65	34.31	28.10	32.21 (4.57)	-						
Rubin-Falcone et al. (2020)		38.43	26.01	29.17	35.99	33.26	27.11	31.66 (4.24)	-						
My approach	Set. 1	**32.79**	**25.23**	**25.71**	**30.87**	**31.20**	**25.14**	**28.49** (3.50)*	29.56	**28.37**	**25.04**	**32.34**	**26.73**	**29.56**	**28.60** (2.53)*
	Set. 2	38.74	26.57	28.56	34.49	31.48	26.23	31.01 (4.91)*	**29.31**	29.82	26.22	32.61	27.61	31.66	29.54 (2.40)*

[a] Abbreviations: RMSE, Root Mean Square Error.
[b] Set. 1: exogenous inputs available; Set. 2: exogenous inputs unavailable.

Table 4.3: Ablation study 1: effects of using glucose-related events in the features set.

	Feature		RMSE (mg/dL)												
Subject			*540*	*544*	*552*	*567*	*584*	*596*	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
30 min	Glucose	-	21.95	17.66	16.67	20.53	21.47	17.02	18.47	**18.06**	15.29	**21.31**	17.46	20.98	18.90 (2.23)
	Full	Set. 1	**19.49***	15.75*	15.68*	**18.96***	**19.52***	15.94*	**17.57**	18.34	14.84	21.69	16.02*	20.08	**17.82 (2.17)**
		Set. 2	20.24*	**15.26***	**15.49***	19.72	20.23*	**15.90***	17.59	18.34	**14.64**	**21.31**	**15.93***	**20.03**	17.89 (2.37)
60 min	Glucose	-	38.96	29.81	29.40	36.14	35.01	28.07	31.59	28.99	25.98	32.93	29.02	32.32	31.52 (3.74)
	Full	Set. 1	**32.79***	**25.23***	**25.71***	**30.87***	**31.20***	**25.14***	29.56*	**28.37**	**25.04**	**32.34**	**26.73***	**29.56***	**28.54 (2.91)***
		Set. 2	38.74	26.57*	28.56	34.49	31.48*	26.23*	**29.31***	29.82	26.22	32.61	27.61	31.66	30.27 (3.77)

[a] Abbreviations: RMSE, Root Mean Square Error.
[b] Set. 1: exogenous inputs available; Set. 2: exogenous inputs unavailable.

Table 4.4: Ablation study 2: effects of using data from other subjects.

	Data		RMSE (mg/dL)												
Subject			*540*	*544*	*552*	*567*	*584*	*596*	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
30 min	Personal	Set. 1	19.97	15.69	16.12	19.05	20.68	16.81	18.93	**18.07**	15.66	21.85	16.60	20.68	18.34 (2.16)
		Set. 2	21.63	16.03	17.00	19.78	20.32	16.79	18.54	18.37	15.50	22.48	16.40	20.53	18.62 (2.32)
	All	Set. 1	**19.49**	15.75	15.68	**18.96**	**19.52***	15.94*	**17.57***	18.34	14.84	21.69	16.02	20.08	**17.82 (2.17)**
		Set. 2	20.24*	**15.26***	**15.49***	19.72	20.23	**15.90***	17.59	18.34	**14.64***	**21.31**	**15.93**	**20.03**	17.89 (2.37)
60 min	Personal	Set. 1	34.68	26.03	26.16	31.44	31.94	27.80	29.53	**27.98**	26.45	34.07	27.56	31.07	29.56 (3.03)
		Set. 2	37.79	26.06	28.42	34.08	32.47	27.42	30.60	30.39	26.36	33.82	28.39	31.49	30.61 (3.53)
	All	Set. 1	**32.79***	**25.23**	**25.71**	**30.87**	**31.20**	**25.14***	29.56	28.37	**25.04***	**32.34**	**26.73**	**29.56***	**28.54 (2.91)***
		Set. 2	38.74	26.57	28.56	34.49	31.48	26.23	**29.31**	29.82	26.22	32.61	27.61	31.66	30.27 (3.77)

[a] Abbreviations: RMSE, Root Mean Square Error.
[b] Set. 1: exogenous inputs available; Set. 2: exogenous inputs unavailable.

Chapter 5

Meta-optimised Time-index Modeling for Glucose Excursion Prediction

This chapter has been published as

R. Cui, H. Suominen, C. J. Nolan and E. Daskalaki (2024). 'Meta-optimised Time-index Modeling for Glucose Excursion Prediction'. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 1911–1918

The corresponding work is open-sourced at https://github.com/r-cui/MOTIMGluPred.

Author contributions: R.C., H.S., C.J.N, and E.D. conceptualised the study; R.C. designed and performed research; R.C. wrote the manuscript and revised it, based on critical comments by H.S., C.J.N., and E.D.

This study related to all the three thesis objectives, as introduced in Section 1.2, namely "modeling", "adaptation" and "engineering". After the last study, I moved my research outside of the scope of historical-value model based glucose prediction, in which researchers have mostly been focusing on proposing novel model designs that improve the predictive performance. A bottleneck of the historical-value modeling paradigm is that there exists an intrinsic limitation on the efficiency as a complex sequence-level mapping must be learned by the historical-value model. With the current trend of deploying healthcare applications on smart devices (Daskalaki et al., 2022), the efficiency of the algorithm becomes another key concern from an engineering perspective, especially in data-driven glucose prediction where deep models are widely used. However, as to be shown later in this chapter, current deep historical-value models in data-driven glucose prediction tend to be heavy in the amount of computation and the model size, limiting its potential in deployment to real-life application. To this end, I turned to investigate how time-index modeling, as an alternative to historical-value modeling, can be applied to the problem of glucose prediction. Compared to historical-value modeling, time-index based modeling was highly more efficient as the learned mapping was at value-level. In this study, I applied the recent advancement of the MOTIM (Woo et al., 2023) to the problem of glucose excursion prediction, and also investigated the performance of the MOTIM model in instance adaptation. In the experiments using two datasets, OhioT1DM and UMT1DM, the MOTIM achieved performance comparable to state-of-the-art historical-value models, e.g., 1-hour excursion average MAPE of 10.25 on OhioT1DM and 12.50 on UMT1DM, while being significantly more efficient in the implementation, e.g., only 1.3% model parameters were required resulting in more than 10 times faster inference. In the

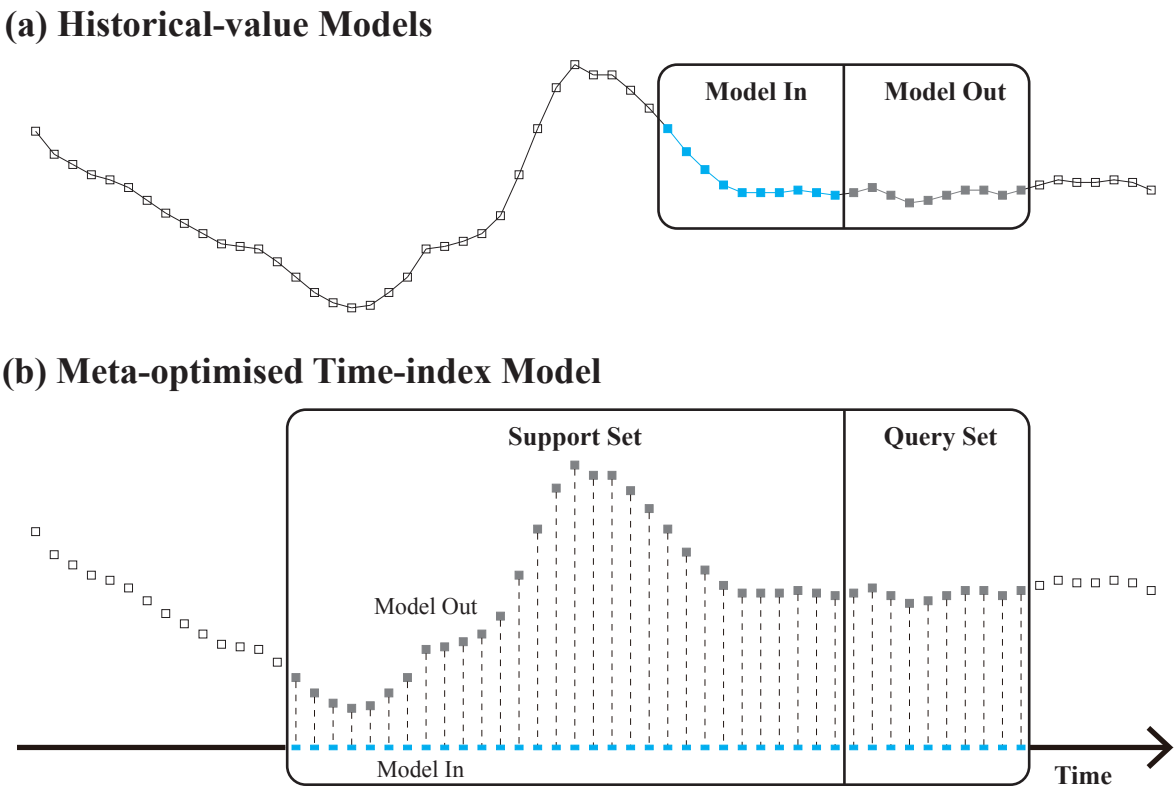


Figure 5.1: Illustration of the key difference between historical-value models and the MOTIM in predicting glucose excursion. Historical-value models take a sequence of past glucose as the input and predict sequence-level output. The MOTIM discretely takes an individual time-index feature as the input and outputs the corresponding glucose value, and it refers to a lookback window to learn the inductive bias. A larger lookback window is marked in the MOTIM diagram to make distinction to the shorter input length in the historical-value models. Figure is inspired by Woo et al. (2023) and adapted from Cui, Suominen et al. (2024). Abbreviations: MOTIM, Meta-optimised Time-index Model.

comparison experiments of instance adaptation between the MOTIM and existing historical-value models, the MOTIM also achieved comparable predictive performance. These results indicated the effectiveness of applying the MOTIM to data-driven glucose prediction as a novel methodology, alternative to the existing historical-value based paradigm. This model not only drastically increased the algorithmic efficiency, benefiting its deployment to smart devices from an engineering perspective, but also could open new research direction of the time-index-based paradigm.

5.1 Introduction

Glucose prediction, as an augmented function of a standard CGM mobile application, would ideally run on the same smart device that the CGM device binds with, e.g., the Guardian system of Medtronic CGM (Medtronic, 2024). In this consideration, the efficiency and engineering of the prediction algorithm is of essential importance, especially in data-driven glucose prediction where deep models are often used. The widely investigated historical-value modeling (Felizardo et al., 2021) in data-driven glucose prediction has the intrinsic bottleneck in this aspect, since all historical-value models learn a mapping from recent glucose sequences to the target future values. As illustrated in Figure 5.1(a), to predict a future glucose

value, such as $\hat{g}_{t+1}$, a mapping of the following form is defined by a historical-value model $f(\cdot)$

$$\hat{g}_{t+1} = f([g_t, g_{t-1}, \dots, g_{t-I+1}]) \tag{5.1}$$

where t denotes the current time index and I denotes the length of the input glucose sequence. The complexity of the sequence-level mapping to be learned is positively correlated to the sequence length. To this end, most existing studies using historical-value models choose to set a relative short input sequence, such as 1 hour or 2 hours, meaning 12 data points or 24 data points under a standard sampling interval of 5 minutes. Although the more recent glucose sequence is more informative in predicting the future values, the infeasibility of learning from older glucose sequences could be a bottleneck of further research.

As reviewed in Oviedo, Vehí et al. (2017); Felizardo et al. (2021) and Section 2.1, historical-value model based personalised glucose prediction has yielded promising outcomes. Following this, a line of research has further investigated instance adaptation, which targets aiding the learning of a personalised model via learning from data of other patients, as reviewed in Section 2.2.2. However, regardless of the task being the standard personalised learning or instance adaptation, historical-value models need to learn the distribution of the target glucose values conditioned on the input sequence in $\mathbb{R}^I$. In this study, I argued that this high dimensional sequence-level conditional distribution could cause an intrinsic learning difficulty, especially in instance adaptation where the input sequence distribution $p([g_t, g_{t-1}, \dots, g_{t-I+1})]$ would naturally shift. This limitation in adaptation presents an additional challenge for the historical-value modeling scheme.

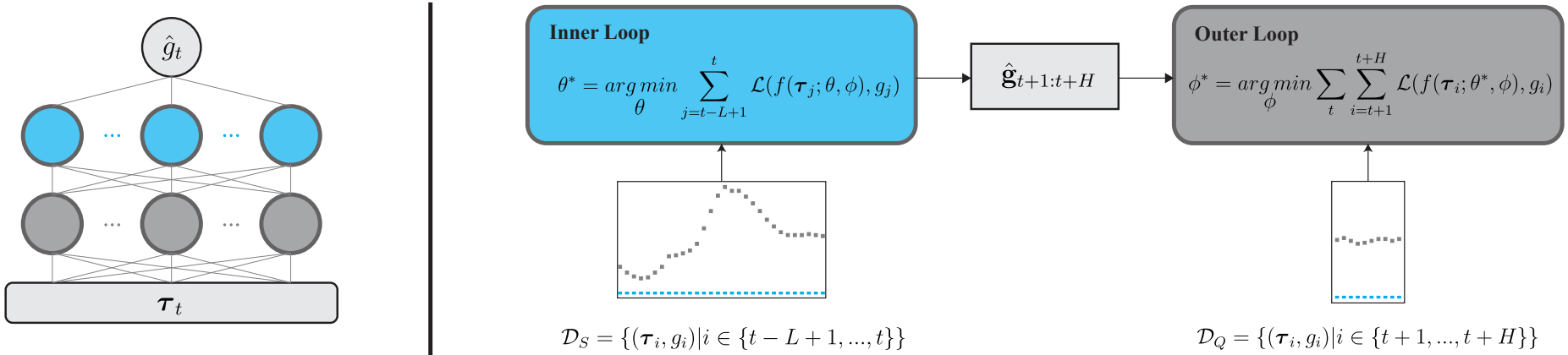
This study hypothesised that such bottlenecks can be largely alleviated with the use of time-index modeling. Recently, the advancement in DL for generic time-series forecasting has proposed the MOTIM (Woo et al., 2023). Different to historical-value models, the MOTIM employs a deep time-index model (Godfrey and Gashler, 2017), which takes a time-index feature as its input and outputs the corresponding glucose value at that time-index, as illustrated by the vertical connections in Figure 5.1(b). More formally, it learns the following mapping:

$$\hat{g}_t = f(\boldsymbol{\tau}_t)$$

where $\boldsymbol{\tau}_t$ is the time-index feature of a specific time-index t. Instead of learning the distribution $p(g_{t+1}|[g_t, g_{t-1}, \dots, g_{t-I+1}])$, the MOTIM learns $p(g_t|\boldsymbol{\tau}_t)$, thus the learning complexity is reduced from a sequence-level to a value-level. However, such change from a horizontal mapping (Figure 5.1(a)) to a vertical mapping (Figure 5.1(a)) makes a sole time-index model limited in making predictions to the future. Thus, the MOTIM uses its meta-optimisation framework to tune the time-index model closer to the local inductive bias before the extrapolation. To this end, the MOTIM retains the capacity of predicting the future sequence by examining the past sequence, akin to historical-value models, while the modeling is purely time-index based which could decouple the learning of the sequence-level mapping in historical-value models.

The objective of this research was to validate if the MOTIM could be a novel solution of data-driven glucose prediction that achieves both high efficiency and predictive performance for both personalised training and instance adaptation. Moreover, I targeted glucose excursion prediction, which required predicting the full trajectory and different to the more common single-point glucose prediction as in Chapter 4. Although being more challenging, the full predicted excursion would be more informative in guiding the patient's decision making. I evaluated the MOTIM in comparison to existing historical-value models in both personalised

training and instance adaptation scenarios on the task of glucose excursion prediction using the OhioT1DM dataset (Marling and Bunescu, 2020) and the UMT1DM dataset (Fox et al., 2018). The MOTIM showed a predictive performance comparable to the state of the art (e.g., 1-hour excursion averaged mean absolute percentage error of 10.25 on OhioT1DM and 12.50 on UMT1DM, and 30-minute RMSE of 20.24 mg/dL on OhioT1DM) while taking a substantially lower cost (e.g., only 1.3% model parameters and more than 10 times faster inference). These results indicated that the MOTIM is a valid methodology and time-index modeling is a promising new direction in the field of data-driven glucose prediction.



(a) Time-index model.

(b) Meta-optimisation framework.

Figure 5.2: (Left) The time-index model architecture, which contains *K*-layered INR (grey nodes) to encode the deep time features and ridge regressor (blue nodes) to generate the predicted glucose values. (Right) The meta-optimisation framework for the time-index model. I make use of blue and grey nodes to mark the correspondences of the group of parameters in the model to be optimised in different stages of the meta-optimisation. Figure is inspired by Woo et al. (2023) and adapted from Cui, Suominen et al. (2024). Abbreviations: INR, Implicit Neural Representation.

5.2 Methodology

In this section, I first formally define the problem of this study, i.e., the glucose excursion prediction, in Section 5.2.1. The MOTIM consisted of two parts that work together, namely the deep time-index model and the meta-optimisation framework, which are introduced respectively in Section 5.2.2 and 5.2.3.

5.2.1 Problem Definition

In this study, the problem was established as a uni-variate forecasting problem following Zarkogianni et al. (2015); Fox et al. (2018); Alfian et al. (2020); Martinsson, Schliep, Eliasson and Mogren (2020) without including exogenous modalities such as administered insulin and carbohydrate information for simplicity as there has not been a consensus of how to incorporate such information into a deep time index model (Godfrey and Gashler, 2017). The aim of glucose excursion prediction was to predict the full trajectory of glucose readings within a prediction horizon. To better illustrate the idea of time-index model for this problem, time-index based subscriptions in the math notations are used in this chapter, which is different to the notations used in Section 4.2.1 which discussed single-point glucose prediction. More discussion on this will be presented in Chapter 7.2.2 which will investigate the difference between single-point prediction and excursion prediction in detail. By indexing the current time as t, the target of glucose excursion prediction was to obtain a predictive model $f(\cdot)$, which is able to predict all the future glucose values $\mathbf{g}_{t+1:t+H} = [g_{t+1}, g_{t+2}, \ldots, g_{t+H}]$ as a sequence. This expanded prediction target, though being more challenging, could provide more information to aid the decision making of the stakeholders when deployed in real-life, and could be helpful in building the trustworthiness of the predictive algorithm to the stakeholders.

The applied methodology MOTIM (Woo et al., 2023) to this problem consisted of two main parts, the time-index model and the meta-optimisation framework. The technical details, derived from (Woo et al., 2023), are to be presented next.

5.2.2 Deep Time-index Model

Different to a historical-value modeling as in Equation 5.1 and Chapter 4, the deep time-index model in the MOTIM, as shown in Figure 5.2(a), learned the following mapping:

$$f(\boldsymbol{\tau}_t) = \hat{g}_t$$

where $\boldsymbol{\tau}_t$ denoted the time-index feature of a specific time-index t, and $\hat{g}_t$ was the predicted glucose value at time t. It took the Gaussian Fourier features augmented relative coordinates as the time-index feature (Tancik et al., 2020). More specifically, for a one-dimensional coordinates vector $\mathbf{c}$ of length $L + H$ where L denoted the length of the lookback window, the high-dimensional time-index feature $\boldsymbol{\tau}$ was given by

$$\gamma(\mathbf{c}) = [\sin(2\pi\mathbf{B}\mathbf{c}), \cos(2\pi\mathbf{B}\mathbf{c})] \in \mathbb{R}^{(L+H)\times D}.$$

Here, D denoted the hidden dimension of the model, $\mathbf{B} \in \mathbb{R}^{D/2}$ was a non-trainable parameter sampled from Gaussian distribution $\mathcal{N}(0, \sigma^2)$ scaled with hyperparameter σ^2, and the coordinates vector $\mathbf{c}$ was a $[0, 1]$ normalised time-indices vector given by

$$\mathbf{c}_{t+i} = \frac{i + L}{L + H - 1}$$

Table 5.1: Conceptual comparison between meta-learning with historical-value models and the meta-optimisation framework of MOTIM.

	Historical-value Model	MOTIM
Modeling	$f(\mathbf{g}_{t-I+1:t}) = \mathbf{g}_{t+1:t+H}$	$f(\boldsymbol{\tau}_t) = \hat{g}_t$
Task	existing patients $\rightarrow$ new patient	$\mathbf{g}_{t-L+1:t} \rightarrow \mathbf{g}_{t+1:t+H}$
Support Set	existing patients	$\{(\boldsymbol{\tau}_i, g_i) \vert i \in \{t-L+1, \ldots, t\}\}$
Query Set	new patient	$\{(\boldsymbol{\tau}_i, g_i) \vert i \in \{t+1, \ldots, t+H\}\}$

[a] Abbreviations: MOTIM, Meta-optimised Time-Index Model.

for $i = -L, \ldots, H-1$. To this end, at each position in the window $L+H$, its time feature became a concatenation of $D/2$ Fourier features with different frequencies, thus ensured a sufficiently expressive yet distinguishable representation across all time-indices.

As illustrated in Figure 5.2(a), the time-index feature $\boldsymbol{\tau}_t$ was then encoded by the implicit neural representations (INR) module (Sitzmann et al., 2020), which was a K-layered, rectified linear unit (ReLU) (Nair and Hinton, 2010) activated fully connected network, to learn the high dimensional time features. This encoding was given by

$$\begin{aligned} \mathbf{z}^{(0)} &= \boldsymbol{\tau} \\ \mathbf{z}^{(k+1)} &= \max(0, \mathbf{z}^{(k)}\mathbf{W}^{(k)} + \mathbf{b}^{(k)}) \end{aligned}$$

in which $\mathbf{W}^{(k)} \in \mathbb{R}^{D\times D}$, $\mathbf{b}^{(k)} \in \mathbb{R}^{D}$, and $\mathbf{z}^{(k)} \in \mathbb{R}^{(L+H)\times D}$ for k in $0, \ldots, K-1$ constituted the trainable parameters of the INR. The final output of the INR, $\mathbf{z}^{(K)}$, was the learned position-wise representation. To produce the predicted glucose values, a final ridge regressor layer (Bertinetto et al., 2018) was subsequently applied, given by

$$\hat{\mathbf{g}} = \mathbf{z}^{(K)}\mathbf{w}_R + b_R \tag{5.2}$$

where $\mathbf{w}_R \in \mathbb{R}^D$ and $b_R \in \mathbb{R}$ were the trainable parameters. The fully linear mapping based deep time-index network provided architectural assurance for the efficiency of the MOTIM.

5.2.3 Meta-optimisation Framework

As mentioned in Section 2.1.1, meta-learning (Hospedales et al., 2021) is the process of adapting a model trained on an original task to a new task, where the *task* refers to a learning problem. The *support set* $\mathcal{D}_S$ and *query set* $\mathcal{D}_Q$, corresponding to the dataset of the original task and the new task, are respectively used in the training process named *inner loop* and *outer loop*. In existing studies using historical-value model that formalise the instance adaptation of glucose prediction with meta-learning concepts as reviewed in Section 2.2.2, learning for existing and unseen patients are defined as the original and new task, respectively. A summary of these concepts is presented in Table 5.1. On this basis, an off-the-shelf pre-trained model using data from existing patients via the outer loop can be quickly adapted to an unseen patient at the cost of an extra inner loop training.

Distinguished from existing practice of adapting a pre-trained historical-value model to an unseen patient via meta-learning, the meta-optimisation framework of the MOTIM enabled the time-index model to make an excursion prediction close to a recent glucose pattern,

otherwise the prediction made via only the extrapolated time-index feature could be weak (Woo et al., 2023). In this scheme, the task was defined as optimising on part of a glucose sequence, rather than optimisation on existing patients as in historical-value models. On this basis, the meta-optimisation framework in the MOTIM was essentially a generalised scope compared to the meta-learning with historical-value models. That is, instance adaptation can be regarded as only a specific application of the MOTIM in which data from multiple patients can be interpreted as a concatenated long glucose sequence. However, data from a single patient which by itself is a long glucose sequence can surely be used to MOTIM. At each time t where a future excursion prediction of length H was desired, the observed examples in the lookback window

$$\mathcal{D}_S = \{(\boldsymbol{\tau}_i, g_i) | i \in \{t-L+1, \ldots, t\}\}$$

was treated as the support set, and the examples in the prediction horizon

$$\mathcal{D}_Q = \{(\boldsymbol{\tau}_i, g_i) | i \in \{t+1, \ldots, t+H\}\}$$

served as the query set. A comparison between the meta-learning concepts in historical-value models and in the MOTIM is summarised in Table 5.1.

To achieve this, the parameters of the full model were divided into two groups

$$\begin{aligned} \theta &= \{\mathbf{w}_R, b_R\} \\ \phi &= \{\mathbf{W}^{(k)}, \mathbf{b}^{(k)} | k \in \{0, \ldots, K-1\}\} \end{aligned}$$

to be optimised in different phases, as distinguished by different colors in Figure 5.2(a). The formulated optimisation target of this bi-level optimisation problem was given by

$$\phi^* = \arg\min_{\phi} \sum_{t} \sum_{i=t+1}^{t+H} \mathcal{L}(f(\boldsymbol{\tau}_i; \theta^*, \phi), g_i) \tag{5.3}$$

$$\text{s.t. } \theta^* = \arg\min_{\theta} \sum_{j=t-L+1}^{t} \mathcal{L}(f(\boldsymbol{\tau}_j; \theta, \phi), g_j), \tag{5.4}$$

where $\mathcal{L}(\cdot, \cdot)$ denoted the loss function, meaning that the outer loop (Equation 5.3) optimised ϕ over all possible windows of length $(L+H)$ from the training data in which in each instance the parameter θ had already been optimised by the inner loop (Equation 5.4). In other words, the training was essentially the learning of the INR representation $\mathbf{z}^{(K)}$ via the outer loop in Equation 5.3, while the fit to local glucose pattern within the lookback window was carried out by the inner loop in Equation 5.4.

The inner loop optimisation targeted only the optimisation of the ridge regressor (Equation 5.2), i.e., find $\mathbf{w}_R^*$ and b_R^* that best fit $\mathbf{g} = \mathbf{z}^{(K)}\mathbf{w}_R + b_R$. Note that this optimisation involved only the inner loop, and both the glucose values $\mathbf{g}$ and the INR output $\mathbf{z}^{(K)}$ were given. Therefore, close-form solution $\mathbf{w}^* = [\mathbf{w}_R^*, b_R^*]$ can be obtained by analytically solving the linear system, given by

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{z}^{(K)}\mathbf{w} - \mathbf{g}\|^2 + \lambda\|\mathbf{w}\|^2 \tag{5.5}$$

$$= (\mathbf{z}^{(K)^T}\mathbf{z}^{(K)} + \lambda\mathbf{I})^{-1}\mathbf{z}^{(K)^T}\mathbf{g}. \tag{5.6}$$

This differentiable close-form solution of inner loop enabled the outer loop to be optimised by regular gradient-based back-propagation learning (Rumelhart, Hinton et al., 1986). An illustration of the full meta-optimisation process is shown in Figure 5.2(b).

5.3 Experiments

This section introduces the experiments evaluating the MOTIM. The experiment setups are described in Section 5.3.1. The evaluation metrics and the compared benchmarks are presented in Section 5.3.2 and 5.3.3, respectively. The results are discussed in Section 5.3.4.[1]

5.3.1 Experiment Setups

I conducted the following experiments to validate the effectiveness of the MOTIM using the OhioT1DM and the UMT1DM datasets. A 60-minute excursion was set as the horizon following the common practice (Oviedo, Vehí et al., 2017). In all experiments, the `mean(std)` of a same set of 12 randomly selected patients from the employed dataset was reported.

Personalised Training Experiment In this experiment, each model was both trained and evaluated using the full training and testing data from the same patient. This experiment corresponded to the standard personalised training and evaluation, which was consistent with most data-driven glucose prediction studies. On this basis, a six-hour segmented experiment was conducted as an extended evaluation to further investigate the performance of MOTIM. In this evaluation, the personalised MOTIM model was tested in different subsets of testing data differing in the time of a day. A day was evenly split into four six-hour segments that represented midnight, morning, afternoon and evening, and the predictive performance on these four subsets of testing data was reported and compared.

Instance Adaptation Experiment In this experiment, one patient was fixed as an anchor and the next patient in the patient list was set as its auxiliary patient. I evaluated first using all training data from the auxiliary patient to pre-train the model to its convergence, then fine-tuning the model using the anchor patient's personal training data. To investigate the adaptation process, I set a hyperparameter *portion of usage* (PoU) controlling the portion of personalised training data to use in the fine-tuning and experimented the performance of *PoU* taking value of $\{0.0, 0.5, 1.0\}$[2].

5.3.2 Evaluation Metrics

To have a comprehensive performance comparison, the glucose excursion prediction performance using the following metrics was evaluated.

Numerical Accuracy As per a generic time-series forecasting task, RMSE and MAPE were employed to evaluate the numerical accuracy. The two metrics at a certain index $i = 1, \ldots, H$ in the prediction horizon were given by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum^{N}(g_{t+i} - \hat{g}_{t+i})^2},$$
$$\text{MAPE} = \frac{1}{N}\sum^{N}\left|\frac{g_{t+i} - \hat{g}_{t+i}}{g_{t+i}}\right|,$$

where N denoted the total number of testing examples. Note that the RMSE was the same metric to the previous study as in Section 4.3.1, yet generalised to the excursion prediction case. The reported scores were averaged over all positions in the predicted excursion. I

[1] Dataset details can be found in Appendix A. Implementation details can be found in Appendix B.
[2] More information can be found by referring to the `self_train_subset_factor` in the codes.

also evaluated the single-point numerical accuracy at the 30$^{\text{th}}$ minute and 60$^{\text{th}}$ minute to enable comparison with the numerous existing studies that conducted single-point prediction, including the study in Chapter 4.

Trend Polarity Differing from the RMSE and MAPE, which only captured the absolute distance of a predicted value against the true value, the trend polarity was also evaluated, which reflected the prediction quality in the increasing or decreasing trend. To mitigate the effect of local noise due to the short sampling interval (5 minutes), I designed the following smoothed trend polarity at a timestamp t in a predicted excursion which included a tolerance factor δ to capture the overall polarity in a time range of $(2\delta + 1)$ samples, given by

$$\text{TP}_t = \begin{cases} 1, & (\sum_t^{t+\delta} g_t - \sum_{t-\delta}^{t} g_t) > 0; \\ 0, & \text{otherwise.} \end{cases}$$

On this basis, the predicted polarity $\hat{\text{TP}}_t$ was then evaluated together with the ground truth TP_t following a standard classification setup, and the accuracy was adopted as the reported metric.

Abnormal Glycemia Due to patients' efforts to avoid abnormal glycemia in their daily lives, especially hypoglycemia, the onsets of such events in the data were rare, even in some of the experiments there was no presence of abnormal glycemia at all. To this end, I did not follow the traditional clinical research in abnormal glycemia prediction that evaluates the onsets of such events using metrics like sensitivity, specificity and false alarm rate as they tend to achieve extreme values and present high variance. Instead, this performance was measured by if the individual values in the predicted excursion fell into abnormal glycemia or not. More specifically, all individual values in the predicted and the ground truth excursion were filtered into binary labels using the hyperglycemia ($\geq$ 180 mg/dL) and hypoglycemia ($\leq$ 70 mg/dL) thresholds. Then the MCC (Chicco and Jurman, 2020) which took value in $[-1, 1]$ was applied as the metric following previous studies (Oviedo, Contreras et al., 2019; Vehí et al., 2020). In this evaluation, I excluded any patient from calculating the reported scores if and only if the following conditions held for the patient: 1) the ground truth was all negative, and 2) the model correctly predicted them. Under such circumstances, no MCC can be calculated as there did not exist two classes.

Efficiency Apart from the predictive performance, I also evaluated the efficiency of the methodologies from the engineering perspective, which was considered important when running the model on smart phones or low-power devices. In particular, the following metrics were measured: number of MAC operations in model inference, number of model parameters, model inference time, and GPU memory consumption. As this performance was determined by the nature of the methodology, results were reported separate to the main results.

5.3.3 Benchmarks

To enable comparison to the state of the art, I included a naive *Repeat* baseline, a Linear baseline and several existing benchmarks from literature differing in the approach of input sequence encoding as the benchmarks. Note that except for the naive *Repeat*, all methods were in the category of historical-value models. The *Repeat* baseline simply repeated the last known glucose value as the predicted excursion, thus no learning was involved at all. I included the Linear baseline to enable comparison with the MOTIM whose backbone

was also based on linear connections. More specifically, Linear learned a linear mapping $f(\mathbf{g}_{t-I+1:t}) = \mathbf{W}\mathbf{g}_{t-I+1:t} + \mathbf{b}$ to directly map the past glucose values to the predicted excursion, where $\mathbf{W} \in \mathbb{R}^{H \times I}$ and $\mathbf{b} \in \mathbb{R}^{H}$ were its trainable parameters.

The three benchmarks from literature differed in the way of encoding the input glucose sequence $\mathbf{g}_{t-I+1:t}$ into a deep feature, namely recurrent, convolutional, and self-attention based modeling. First, for recurrent models (Fox et al., 2018; Sun et al., 2018; Mirshekarian et al., 2019), I included the multi-output GRU (MoGRU) (Fox et al., 2018) into the comparison, which took the deep feature $\mathbb{R}^{D}$ and applied H individual linear mappings $\mathbb{R}^{D} \rightarrow \mathbb{R}$ to predict the H values in $\hat{\mathbf{g}}_{t+1:t+H}$ all at once. Second, for CNN based models, the sequence-to-sequence convolution (SeqConv) (Sutskever et al., 2014) method as studied by (Zaidi et al., 2021) in the task of glucose prediction was adopted, where a dilated convolution based multi-layer encoder and decoder block were respectively employed to encode the past glucose sequence to a deep feature $\mathbf{h} \in \mathbb{R}^{I \times D}$ and decode $\mathbf{h}$ to the values in the prediction one at a time. Third, for self-attention (Vaswani et al., 2017) based methods, the RSAN as in Chapter 4 was included into the comparison. Note that these benchmarks were different to those in Chapter 4 in order to ensure the coverage of different encoding backbones, namely RNN, CNN, and Transformer.

Table 5.2: Personalised training experiment results of the MOTIM and historical-value benchmarks. Scores are reported in `mean(std)`, and results with best mean value are marked as bold. The MOTIM results that were significantly better ($p < 0.05$) than at least half of the benchmarks are underlined.

	Model	RMSE↓			MAPE↓			TP ACC↑		Abnormal glycemia MCC↑	
		all	at 30 min	at 60 min	all	at 30 min	at 60 min	δ=1	δ=2	Hyperglycemia	Hypoglycemia
OhioT1DM	*Repeat*	25.73 (2.98)	23.28 (2.74)	38.16 (4.45)	11.64 (2.03)	11.35 (1.98)	19.29 (3.44)	-	-	74.83 (5.10)	43.67 (19.61)
	Linear	24.38 (2.61)	22.21 (2.75)	35.62 (3.87)	11.54 (2.06)	11.13 (2.05)	18.89 (3.63)	57.82 (2.11)	59.67 (2.10)	75.17 (4.85)	37.06 (13.26)
	MoGRU	22.64 (2.45)	**20.17 (2.42)**	33.78 (3.55)	10.55 (2.12)	10.08 (2.07)	17.56 (3.32)	**60.78 (2.59)**	**61.95 (2.68)**	**78.45 (4.45)**	33.90 (17.41)
	RSAN	24.87 (3.14)	22.63 (3.05)	35.76 (4.15)	11.99 (2.58)	11.45 (2.53)	18.71 (3.85)	60.31 (3.26)	61.31 (3.29)	75.41 (4.67)	24.85 (17.64)
	SeqConv	25.95 (2.35)	23.65 (2.27)	37.25 (3.33)	13.10 (2.35)	12.50 (2.21)	20.54 (3.75)	59.71 (2.91)	60.30 (2.97)	73.08 (6.30)	16.74 (13.19)
	MOTIM	**22.44 (2.75)**	20.24 (2.83)	**33.17 (3.80)**	**10.25 (1.81)**	**9.86 (1.76)**	**17.28 (3.23)**	60.06 (2.85)	61.05 (2.97)	78.19 (4.10)	**41.54 (15.35)**
UMT1DM	*Repeat*	28.13 (6.07)	25.93 (5.64)	40.64 (9.19)	12.79 (4.03)	12.58 (4.01)	20.30 (6.45)	-	-	72.73 (9.30)	50.26 (25.43)
	Linear	26.41 (5.72)	24.79 (5.71)	37.20 (8.34)	13.64 (4.76)	13.33 (4.64)	21.65 (7.76)	53.89 (3.30)	56.56 (3.20)	73.93 (10.35)	**44.76 (18.18)**
	MoGRU	25.68 (5.89)	23.77 (5.71)	36.55 (8.36)	12.86 (4.67)	12.43 (4.69)	20.89 (7.76)	55.00 (4.16)	56.73 (3.45)	**74.22 (9.64)**	43.41 (23.19)
	RSAN	27.57 (5.76)	25.57 (5.48)	38.46 (8.44)	14.36 (5.87)	13.84 (5.76)	21.24 (7.88)	**56.25 (3.73)**	57.48 (3.73)	73.20 (10.26)	39.47 (25.13)
	SeqConv	28.79 (6.66)	26.73 (6.17)	40.27 (9.68)	15.97 (6.73)	15.40 (6.52)	24.17 (9.70)	54.38 (4.59)	55.46 (4.10)	70.85 (8.53)	28.22 (16.96)
	MOTIM	**25.31 (5.68)**	**23.62 (5.78)**	**35.75 (8.06)**	**12.50 (4.71)**	**12.09 (4.68)**	**20.23 (7.92)**	54.82 (4.45)	**56.90 (4.20)**	73.39 (15.93)	44.14 (25.49)

[a] Abbreviations: RMSE, Root Mean Square Error; MAPE, Mean Absolute Percentage Error; TP, Trend Polarity; ACC, Accuracy; MCC, Matthews's Correlation Coefficient; MoGRU, Multi-output Gated Recurrent Unit; RSAN, Recurrent Self-attention; SeqConv, Seq-to-seq Convolution; MOTIM, Meta-Optimised Time-index Model.

[b] The ACC and MCC scores are presented in percentages to better show the close differences. Thus, ACC takes value in range $[0, 100]$ and MCC takes value in range $[-100, 100]$.

Table 5.3: Instance adaptation experiments results of the MOTIM and historical-value benchmarks. Scores are reported in `mean(std)`, and results with best mean value are marked as bold. The MOTIM results that were significantly better ($p < 0.05$) than at least half of the benchmarks are underlined.

	PoU	Model	RMSE↓			MAPE↓			TP ACC↑		Abnormal glycemia MCC↑	
			all	at 30 min	at 60 min	all	at 30 min	at 60 min	δ=1	δ=2	Hyperglycemia	Hypoglycemia
OhioT1DM	-	*Repeat*	25.73 (2.98)	23.28 (2.74)	38.16 (4.45)	11.64 (2.03)	11.35 (1.98)	19.29 (3.44)	-	-	74.83 (5.10)	43.67 (19.61)
	0.0	Linear	24.57 (2.52)	22.67 (2.73)	35.90 (3.84)	11.75 (1.91)	11.47 (1.88)	19.33 (3.34)	57.39 (1.86)	59.09 (2.13)	75.34 (5.19)	37.45 (13.84)
		MoGRU	23.04 (2.61)	20.47 (2.63)	34.48 (3.73)	10.87 (1.88)	10.29 (1.90)	18.45 (3.19)	**60.33 (2.34)**	**61.01 (2.52)**	77.86 (5.56)	36.40 (13.90)
		RSAN	25.45 (3.40)	23.08 (3.20)	36.65 (4.65)	12.53 (2.34)	11.90 (2.25)	19.62 (3.58)	58.36 (3.02)	59.62 (2.85)	74.33 (5.53)	21.72 (17.19)
		SeqConv	27.06 (3.03)	24.66 (2.90)	38.78 (4.14)	14.15 (2.60)	13.51 (2.49)	22.16 (4.21)	58.37 (2.42)	58.91 (2.11)	73.13 (6.32)	13.69 (12.27)
		MOTIM	**22.71 (2.64)**	**20.38 (2.71)**	**33.71 (3.78)**	**10.25 (1.77)**	**9.75 (1.62)**	**17.55 (3.38)**	59.88 (2.38)	60.71 (2.44)	**78.52 (4.08)**	**43.65 (14.56)**
	0.5	Linear	23.84 (2.54)	21.46 (2.28)	35.17 (3.95)	11.19 (2.03)	10.71 (1.84)	18.74 (3.61)	58.05 (2.25)	59.45 (2.60)	76.08 (4.50)	39.97 (12.44)
		MoGRU	**22.57 (2.49)**	**20.14 (2.45)**	33.64 (3.60)	10.49 (2.18)	9.99 (2.12)	17.55 (3.44)	**61.12 (2.41)**	**62.26 (2.74)**	**78.67 (4.51)**	36.00 (18.38)
		RSAN	25.40 (2.56)	23.09 (2.47)	36.44 (3.71)	12.49 (2.15)	11.86 (2.15)	19.35 (3.32)	59.74 (2.76)	60.91 (2.68)	75.14 (4.54)	17.84 (19.30)
		SeqConv	26.09 (2.33)	23.86 (2.26)	37.24 (3.33)	13.38 (2.35)	12.81 (2.30)	20.82 (3.74)	59.12 (2.39)	59.55 (2.22)	72.53 (7.74)	16.08 (16.11)
		MOTIM	22.66 (2.83)	20.51 (2.96)	**33.43 (3.85)**	**10.32 (1.85)**	**9.93 (1.70)**	**17.34 (3.26)**	60.13 (3.01)	61.10 (3.09)	78.07 (3.75)	**42.61 (13.05)**
	1.0	Linear	23.34 (2.58)	20.95 (2.41)	34.72 (3.96)	10.86 (2.04)	10.45 (1.87)	18.42 (3.64)	58.83 (2.24)	59.90 (2.33)	76.89 (4.61)	41.23 (12.04)
		MoGRU	22.58 (2.51)	**20.06 (2.26)**	33.79 (3.77)	10.55 (2.10)	10.01 (1.93)	17.87 (3.56)	60.59 (3.08)	61.79 (3.25)	**78.66 (4.73)**	35.88 (18.16)
		RSAN	24.88 (2.87)	22.61 (2.80)	35.77 (3.90)	11.94 (2.45)	11.36 (2.37)	18.56 (3.64)	**60.64 (2.92)**	**61.98 (3.16)**	74.51 (5.83)	19.99 (15.53)
		SeqConv	25.17 (3.32)	22.90 (3.17)	36.39 (4.61)	12.26 (2.50)	11.77 (2.43)	19.34 (3.80)	60.44 (2.81)	60.98 (2.91)	74.29 (5.79)	19.39 (14.39)
		MOTIM	**22.46 (2.51)**	20.26 (2.55)	**33.22 (3.66)**	**10.24 (1.71)**	**9.85 (1.57)**	**17.27 (3.16)**	59.94 (2.48)	61.00 (2.70)	78.07 (3.93)	**41.94 (13.41)**

	PoU	Model	RMSE↓			MAPE↓			TP ACC↑		Abnormal glycemia MCC↑	
			all	at 30 min	at 60 min	all	at 30 min	at 60 min	δ=1	δ=2	Hyperglycemia	Hypoglycemia
	-	*Repeat*	28.13 (6.07)	25.93 (5.64)	40.64 (9.19)	12.79 (4.03)	12.58 (4.01)	20.30 (6.45)	-	-	72.73 (9.30)	50.26 (25.43)
UMT1DM	0.0	Linear	27.08 (5.95)	25.28 (5.71)	37.77 (8.31)	13.61 (4.55)	13.22 (4.41)	21.20 (7.60)	53.64 (3.25)	56.14 (3.08)	73.87 (10.57)	45.43 (22.75)
		MoGRU	25.80 (6.20)	23.98 (6.20)	36.65 (8.56)	12.43 (3.92)	12.08 (3.81)	**19.87 (6.19)**	55.09 (3.97)	**57.16 (3.93)**	**74.71 (9.62)**	48.55 (22.53)
		RSAN	28.21 (5.90)	26.37 (5.61)	39.23 (8.40)	13.92 (3.80)	13.45 (3.73)	20.84 (5.68)	53.92 (3.93)	55.90 (4.06)	69.79 (12.62)	42.12 (28.48)
		SeqConv	29.33 (6.81)	27.51 (6.28)	40.12 (9.67)	16.33 (5.72)	15.73 (5.37)	24.21 (9.01)	53.88 (3.90)	55.89 (2.92)	70.17 (8.75)	27.95 (22.04)
		MOTIM	**25.40 (6.18)**	**23.47 (6.14)**	**36.16 (9.05)**	**12.41 (4.59)**	**12.03 (4.52)**	19.96 (7.70)	**55.24 (4.65)**	57.09 (4.73)	73.04 (19.17)	**52.51 (24.32)**
	0.5	Linear	26.03 (5.69)	24.29 (5.59)	36.87 (8.22)	13.33 (4.82)	13.03 (4.73)	21.42 (7.83)	54.62 (4.11)	56.82 (3.61)	74.57 (10.21)	46.11 (17.65)
		MoGRU	25.68 (5.95)	23.93 (5.92)	36.52 (8.62)	**12.54 (4.29)**	**12.11 (4.17)**	20.41 (7.20)	54.99 (3.73)	56.71 (3.84)	73.64 (10.00)	**50.67 (22.04)**
		RSAN	27.37 (6.44)	25.44 (6.14)	38.15 (8.95)	13.99 (4.64)	13.42 (4.53)	21.07 (7.12)	55.14 (4.77)	56.99 (4.68)	71.15 (10.12)	40.63 (21.31)
		SeqConv	28.79 (6.82)	26.88 (6.24)	39.92 (10.09)	16.68 (7.10)	16.13 (6.89)	25.05 (10.72)	**55.23 (4.07)**	57.09 (2.90)	72.37 (11.05)	23.56 (17.38)
		MOTIM	**25.55 (5.73)**	**23.76 (5.78)**	**36.22 (8.09)**	12.60 (4.40)	12.21 (4.40)	**20.36 (7.17)**	55.06 (3.87)	**57.16 (3.28)**	**75.18 (10.29)**	44.54 (25.51)
	1.0	Linear	25.74 (5.62)	23.94 (5.55)	36.60 (8.13)	13.02 (4.61)	12.70 (4.57)	21.10 (7.55)	54.57 (4.18)	56.74 (3.54)	**74.59 (10.02)**	46.99 (19.01)
		MoGRU	25.34 (5.71)	23.48 (5.66)	36.18 (8.16)	12.58 (4.71)	12.10 (4.75)	20.54 (7.62)	55.75 (4.67)	57.39 (4.31)	73.97 (10.62)	**52.75 (18.62)**
		RSAN	27.30 (5.85)	25.41 (5.54)	38.13 (8.62)	13.77 (4.68)	13.22 (4.54)	20.74 (7.23)	**56.14 (3.24)**	**57.58 (3.21)**	73.30 (9.74)	39.68 (23.75)
		SeqConv	28.31 (6.14)	26.37 (5.80)	39.55 (8.73)	15.59 (5.07)	15.11 (5.08)	23.62 (7.43)	54.57 (4.10)	56.42 (3.67)	73.08 (10.03)	29.49 (16.92)
		MOTIM	**25.21 (5.60)**	**23.47 (5.62)**	**35.61 (7.95)**	**12.57 (4.83)**	**12.03 (4.65)**	**20.48 (8.17)**	54.94 (4.52)	57.07 (4.32)	73.10 (16.33)	43.25 (25.33)

[a] Abbreviations: RMSE, Root Mean Square Error; MAPE, Mean Absolute Percentage Error; TP, Trend Polarity; ACC, Accuracy; MCC, Matthews's Correlation Coefficient; MoGRU, Multi-output Gated Recurrent Unit; RSAN, Recurrent Self-attention; SeqConv, Seq-to-seq Convolution; MOTIM, Meta-Optimised Time-index Model; PoU, Portion of Usage.

Table 5.4: Six-hour segmented experiment results of the personalised trained the MOTIM on the OhioT1DM dataset. Scores are reported in `mean(std)`.

	Time	RMSE↓			MAPE↓			TP ACC↑		Abnormal glycemia MCC↑	
		all	at 30 min	at 60 min	all	at 30 min	at 60 min	$\delta = 1$	$\delta = 2$	Hyperglycemia	Hypoglycemia
Repeat	0 - 5	16.13 (2.71)	14.75 (2.50)	23.65 (4.31)	7.91 (1.87)	7.73 (1.74)	13.06 (3.47)	-	-	84.22 (8.78)	41.35 (27.99)
	6 - 11	26.43 (5.19)	24.03 (4.54)	38.94 (8.10)	12.02 (3.58)	11.59 (3.30)	20.21 (6.58)	-	-	63.53 (12.48)	44.53 (28.65)
	12 - 17	28.49 (4.81)	25.73 (4.52)	42.23 (7.19)	12.97 (2.41)	12.80 (2.45)	20.92 (4.05)	-	-	72.97 (11.09)	23.53 (16.09)
	18 - 23	29.57 (6.09)	26.62 (5.45)	44.01 (9.22)	13.51 (3.27)	13.20 (3.07)	22.38 (5.87)	-	-	67.35 (7.59)	37.03 (9.68)
MOTIM	0 - 5	15.15 (3.04)	13.42 (3.36)	22.78 (4.22)	7.87 (1.81)	7.38 (1.74)	13.96 (3.72)	51.83 (7.28)	51.77 (7.64)	82.75 (11.10)	34.28 (18.73)
	6 - 11	23.49 (4.55)	21.54 (4.52)	34.01 (6.19)	11.04 (3.94)	10.66 (3.88)	18.45 (6.82)	63.65 (7.17)	64.85 (7.55)	66.59 (10.03)	36.28 (33.24)
	12 - 17	25.35 (6.75)	23.05 (5.81)	36.99 (11.47)	11.48 (2.24)	11.53 (1.98)	18.22 (5.31)	58.88 (4.15)	59.74 (4.56)	75.48 (9.63)	30.30 (14.31)
	18 - 23	23.82 (4.97)	21.57 (4.45)	34.93 (7.56)	10.80 (2.60)	10.59 (2.46)	17.55 (4.71)	65.34 (4.99)	67.29 (5.39)	68.92 (12.38)	37.00 (19.75)

[a] Abbreviations: MOTIM, Meta-optimised Time-Index Model; RMSE, Root Mean Square Error; MAPE, Mean Absolute Percentage Error; TP, Trend Polarity; ACC, Accuracy; MCC, Matthews's Correlation Coefficient.

Table 5.5: Comparison of the efficiency between the MOTIM and existing benchmarks. The `mean(std)` of inference time over 1000 runs on different mini-batches is reported.

	MAC	Parameter	Time	GPU Memory
MoGRU	1.824G	2.373M	15.88(0.74)ms	1409MiB
RSAN	5.438G	593.665K	65.72(1.57)ms	1417MiB
SeqConv	4.008G	183.745K	37.98(1.24)ms	1861MiB
MOTIM	**786.432K**	**32.896K**	**2.96(1.76)ms**	**1207MiB**

[a] Abbreviations: MOTIM, Meta-optimised Time-Index Model; MoGRU, Multi-output Gated Recurrent Unit; RSAN, Recurrent Self-attention Network; SeqConv, Seq-to-seq Convolution; MAC, Multiply–accumulate Operation.

5.3.4 Results

In the direct comparison of the MOTIM and existing historical-value models shown in Table 5.2, the MOTIM achieved comparable performance to the benchmarks, reflected by overall best performance on most metrics, though only a small portion achieved statistical significance. Among these metrics, MOTIM was not at the first rank in TP scores with a maximal difference of $\sim 2\%$. This observation suggested further potential of improvement in the MOTIM to match the performance of historical-value models, such as MoGRU and RSAN.

The above observation of the effectiveness of MOTIM was consistent in the instance adaptation experiments in Table 5.3, suggesting the robustness of MOTIM. Apart from that, it was noticeable that the MOTIM converged faster by vertically comparing results in $PoU = [0.0, 0.5, 1.0]$. The performance gap of the two experiments of the MOTIM was smaller than other methods (e.g., less than 1mg/dL overall RMSE decrease (22.71 to 22.46) in the MOTIM *v.s.* around 2mg/dL for SeqConv (27.06 to 25.17) in OhioT1DM). This suggested the effectiveness of the MOTIM easing the learning difficulty by changing the sequence-level distribution learning to value-level, since the MOTIM was shown to rely less on the amount of personalised training data.

In addition to analysing the two tables individually, more remarks can be made via a cross-examination between Table 5.2 and 5.3. First, the Table 5.2 can be compared to the $PoU = 0.0$ experiments in Table 5.3. This corresponded to a comparison that the amount of training data was fixed to 1 patient's amount, while the source of training data was different (personalised data *vs.* data from another patient). A consistently better performance can be observed in the personalised experiment when comparing the two experiments entry by entry. Moreover, the performance gap in this comparison was different among the listed methods. More specifically, the MOTIM achieved smaller performance gaps than existing historical-value models (e.g., the performance gap of the first metric RMSE (all) was 22.71-22.44=0.27 mg/dL while the same entry of SeqConv was 27.06-25.95=1.11 mg/dL). This observation demonstrated that 1) the cross-instance variance indeed existed and caused learning difficulty as the personalised training consistently showed better performance, and 2) the MOTIM was more robust to instance change as the performance gap was narrower than the existing historical-value models.

Second, comparison can also be made between Table 5.2 and the $PoU = 1.0$ experiments of the MOTIM in Table 5.3. In this comparison, both experiments ended with a same training

round using full personalised data, while the instance adaptation was preceded with a pre-training phase using data from a different patient. As a result, they exhibited similar performance (e.g., the MOTIM performance gap of RMSE (all) between the two experiments was only 22.46-22.44=0.02 mg/dL). This indicated the feasibility of putting the MOTIM in a pre-training & fine-tuning pipeline: the pre-training phase using another patient's data did not pollute the following personalised training phase and the two experiments ended up with similar convergence.

As shown in Table 5.5, the MOTIM was demonstrated to be significantly more efficient than other historical-value alternatives in terms of runtime and memory usage. Training the MOTIM consumed at least 200MiB GPU memory less than all other methods. In terms of the model structure, the MOTIM contained less than 20% parameters than the lightest counterpart and only 0.01% MACs in the feedforward computation. Compared to MoGRU, which achieved relatively better predictive performance than other historical-value models, the MOTIM contained only 1.3% the number of the MoGRU's parameters and achieved more than 10 times faster inference. Thus, it was clearly demonstrated that the MOTIM achieved significantly higher efficiency than existing methodologies.

5.4 Discussion

This section summaries the principal results, analyses the limitations of this study, and proposes future work and broader impact in the subsections.

5.4.1 Principal Results

This study introduced the utilisation of the MOTIM for the glucose prediction tasks and evaluated it in both personalised training and instance adaptation scenarios. The MOTIM method yielded predictive performance comparable to the current state of the art (1-hour excursion average MAPE of 10.25 on OhioT1DM and 12.57 on UMT1DM) while being significantly more efficient (around 99% parameters reduction and 10 times faster inference). Therefore, the results indicated the prospective value of employing the time-index modeling as an innovative direction in this field.

5.4.2 Limitations

Due to the large amount of experiments in this study, it was not feasible to run the experiments multiple times using different random seeds (a single round of all the presented experiments took around 2 days on a modern GPU equipped computing environment). However, it was observed that the experimental results were stable to different initialisations of model parameters. Besides, due to the natural variance across patients (Hall et al., 2018), the performance of different patients could have large variance, too. To this end, violin plots was provided in Figure 5.3 to further illustrate the scores distribution which supported the effectiveness of the MOTIM compare to other methodologies. For simplicity, a subset of metrics was included in the plotting and only the results of OhioT1DM were presented. A final limitation that likely contributed to the failure of most neural networks to be superior to using a linear model was the limited amount of complex, heterogeneous data. This observation suggests a limitation of existing deep methods that the observed performance increase may not be fully justified given the associated increase in model complexity and computational cost. This is to be further discussed in Section 8.6.

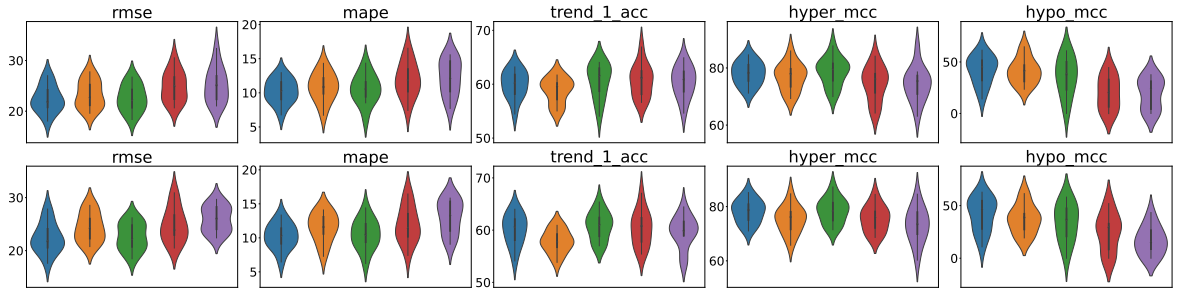


Figure 5.3: Violin plots of the 12-patient results distribution on OhioT1DM dataset. Five benchmarks in each subplot from left to right are: MOTIM, Linear, MoGRU, RSAN, and SeqConv. Two rows are: $PoU = 1.0$ instance adaptation experiment and personalised training experiment. Abbreviations: MOTIM, Meta-optimised Time-Index Model; MoGRU, Multi-output Gated Recurrent Unit; RSAN, Recurrent Self-attention Network; SeqConv, Seq-to-seq Convolution; PoU, Portion of Usage.

5.4.3 Future Work

I hope that the novel modeling based on time-index discussed in this study could bring the research beyond the boundary of traditional historical-value modeling. For example, one essential breakthrough brought by the high efficiency of the MOTIM was that the receptive field of past values was largely increased (Figure 5.1), comparing to historical-value models which focus mostly on the very recent past sequence. A promising future research topic could be using the MOTIM to fit to a longer local pattern (e.g., to a half-day or daily level) which the input sequence would be more heterogeneous, e.g., interpersonal life styles and seasonalities. To achieve this, datasets containing larger personalised data than the employed OhioT1DM and UMT1DM will need to be collected.

5.4.4 Significance and Broader Impact

Time-index modeling is considered to be a novel direction compared to the widely explored historical-value based modeling in doing glucose prediction Oviedo, Vehí et al. (2017); Felizardo et al. (2021). The change of modeling from sequence-level to value-level could open new potentials to this field, such as making use of the longer modeling to learn personalised glucose patterns that cannot be captured by short glucose sequences in purpose of improving the instance adaptation. The notably reduced cost in comparison to prevailing historical-value models would also simplify its implementation on smart devices. I hope that the concepts presented in this research, along with the openly accessible code resources, have the potential to facilitate the research of time-index based glucose prediction to a new stage.

The contribution of this study is listed as follows:

- I proposed for the first time to apply the MOTIM, which is a novel method alternative to the established historical-value based approaches. The method achieved predictive performance comparable to the state of the art.
- I demonstrated that the MOTIM achieved state-of-the-art predictive performance in instance adaptation of glucose excursion prediction.
- I showed the significantly higher efficiency of the MOTIM compared to existing historical-value based glucose prediction models which would facilitate the engineering in real life on smart devices.

Continuing from this study that reached its discussion to the engineering of data-driven glucose prediction, I further studied other aspects of engineering, e.g., from a problem definition perspective. More specifically, I proposed a new strategy for glucose prediction in the postprandial scenario, as the after-meal glucose control is the most essential part of T1DM management. Targeting what could be a good practice of building a predictive model only for postprandial times, the next study is to be presented in Chapter 6.

Chapter 6

Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia

This chapter has been published as

R. Cui, C. J. Nolan, E. Daskalaki and H. Suominen (2023). 'Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes'. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. DOI: 10.1109/EMBC40787.2023.10340094

The corresponding work is open-sourced at
https://github.com/r-cui/PostprandialHyperHypoPrediction.

Author contributions: R.C., C.J.N, E.D., and H.S. conceptualised the study; R.C. designed and performed research; R.C. wrote the manuscript and revised it, based on critical comments by C.J.N., E.D., and H.S.

This study related to the "engineering" objective of the thesis, as introduced in Section 1.2.3. More specifically, this study targeted the prediction of postprandial glucose, which is of great concern in T1DM daily management. Most of existing studies on this topic belong to clinical studies which explore predicting hypoglycemia events after meals using methods like nutrition absorption models (Karim et al., 2020), and less research focus has been devoted to data-driven approaches. This study made the first attempt to establish the problem as the joint prediction of postprandial hyperglycemia and hypoglycemia in a fully data-driven manner. In this study, a strategy for extracting training and testing examples from daily CGM data for this aim was proposed, in which the abnormal glycemia prediction task was defined as a historical-value based binary classification problem. To tackle this, I employed the original LSTM structure as the encoding backbone, which has been shown effective and robust, hence serve the purpose of validating the proposed strategy for extracting training & testing instances. To further improve the engineering quality, a unified LSTM backbone was used to handle the two prediction targets hyperglycemia and hypoglycemia simultaneously via two linear prediction heads, instead of using two individual models. The experiments using the OhioT1DM dataset achieved state-of-the-art performance (MCC of 0.61 for hyperglycemia and 0.48 for hypoglycemia), which exceeded the performance of the selected benchmarks from data-driven and physiological methods and baselines using traditional ML algorithms that I implemented.

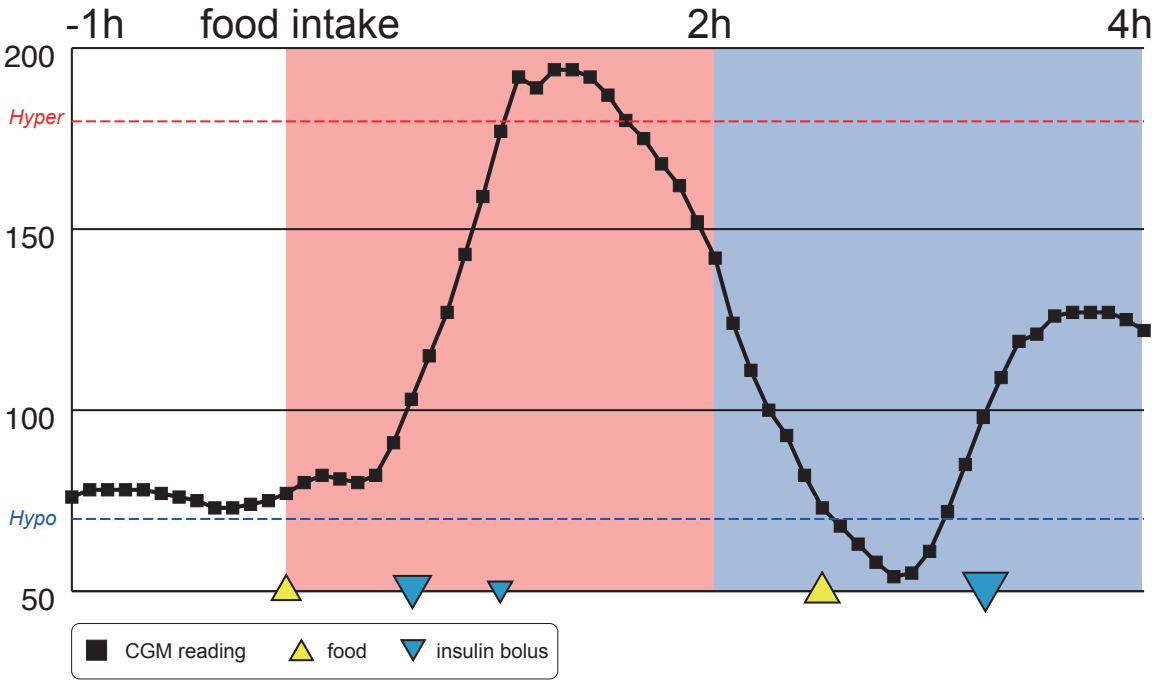


Figure 6.1: A snippet of T1DM patient's real-life CGM data around a meal. This example shows a typical postprandial glucose pattern: the digestion of food causes glucose to peak, followed by its falling to a valley under the effect of natural metabolism and meal-corresponded insulin. Figure is adapted from Cui, Nolan et al. (2023). Abbreviations: CGM, Continous Glucose Monitoring; T1DM, Type 1 Diabetes Mellitus.

6.1 Introduction

A major motivation of this study is that glucose values present different patterns or distribution during a day, while most existing data-driven glucose prediction studies use the full set of CGM data to train the model (Zhu, Li, Herrero, Chen et al., 2018; Martinsson, Schliep, Eliasson and Mogren, 2020; Deng, Lu et al., 2021; Yang, Yu et al., 2022). As shown in Figure 6.1, under the effect of food, postprandial glucose usually has a pattern from rising to declining that differs from other times such as the nocturnal glucose, which is relatively stable. The same phenomenon can also be inferred from Table 5.4 in Chapter 5, as the learning difficulty increased during meal times. To validate this, I conducted a statistical comparison between postprandial 4-hour glucose and glucose at other times using the OhioT1DM dataset to study the distribution difference of postprandial 4 hour glucose values vs. the rest. As shown in Table 6.1, the two-sample Kolmogorov–Smirnov (KS) tests (Massey Jr, 1951) on all patients achieve a p-value less than 0.05, indicating their difference in the distribution. This evidence suggested the potential of treating postprandial glucose prediction separately to all-day prediction, which is the topic of this study.

Prior research has mainly approached postprandial glucose prediction from two directions, namely prediction at meal time (Oviedo, Contreras et al., 2019; Pustozerov et al., 2020; Vehí et al., 2020; Adelberger et al., 2021) and prediction within a fixed horizon (Seo, Lee et al., 2019; Karim et al., 2020). Meal-time prediction considers a pre-defined meal effective time range such as 4 hours after the meal, and all predictions regarding the time range are made at once at the meal time. This scheme reflects the application that the patient could re-plan the meal advised by the prediction result, e.g., adjust the planned food quantity or insulin amount. Although the envisaged application is of great value, a natural disadvantage of this scheme

is the lack of flexibility, i.e., it cannot take factors after the meal intake into consideration. As illustrated in Figure 6.1, other meals and insulin intake events could be present at the patient's wish in real life, rendering the meal-time prediction inaccurate under such circumstances. In contrast, the fixed prediction horizon scheme considers a shorter prediction horizon such as 30 or 60 minutes, and the prediction is rolling starting right after the meal time until it covers all the meal effective time range. This scheme corresponds to the application that once a meal event is announced by the patient, the algorithm constantly runs to predict glucose for a short future, such that later dynamics of the glucose trajectory can be accommodated. Although having a shorter prediction horizon than the meal-time prediction scheme, the application of fixed horizon prediction is more flexible. To this end, this study followed the fixed horizon prediction scheme.

The objective of this study was to design and develop a strategy for the prediction of postprandial abnormal-glycemia events under the fixed horizon prediction scheme. Under this scheme, I proposed a strategy of extracting training and testing examples from daily CGM data in which the prediction target was set to both postprandial hyperglycemia and hypoglycemia instead of only one of them as in previous studies (Oviedo, Contreras et al., 2019; Seo, Lee et al., 2019; Vehí et al., 2020). The two targets were annotated by two distinctive configurations of time zones and two thresholds according to clinical suggestions (Daenen et al., 2010; Altuntaş, 2019). Under this strategy, an LSTM-based model that used the 1-hour past CGM pattern was applied to predict the occurrence of abnormal-glycemia events at a horizon of 30 minutes. In consideration of engineering effectiveness, instead of separately obtaining two models for the two tasks (hypo-, hyper-glycemia), I used a single LSTM backbone to handle the temporal feature extraction for both tasks, thus the need for computational resources in both training and inference was minimised. The proposed methodology was evaluated on the OhioT1DM dataset. Though there existed differences in the data used by different studies, the experiments achieved state-of-the-art results compared to existing literature (MCC of 0.61 for hyperglycemia and 0.48 for hypoglycemia). Moreover, to further study the effect of the aforementioned glucose distribution shift in postprandial scenario, I conducted an ablation experiment in which the model was trained on the whole CGM dataset. The resulting performance deterioration and 10-fold increase of training time indicated that treating postprandial glucose prediction separately was a valid approach that should be further pursued.

6.2 Methodology

The proposed problem definition, i.e., the joint prediction of short-term hyperglycemia and hypoglycemia using daily CGM data, is provided in Section 6.2.1. The proposed methodology is presented in Section 6.2.2 and 6.2.3.

Table 6.1: Distribution in the format of mean(std) of postprandial 4-hour CGM data and other CGM data in the OhioT1DM dataset, and two-sample KS tests per patient.

Patient ID	540	544	552	584	596	559	563	570	575	588	591
Postprandial 4h CGM	139.7 (60.7)	183.5 (59.5)	145.3 (55.7)	186.1 (64.2)	147.9 (51.1)	173.8 (72.8)	149.8 (51.3)	181.1 (66.2)	137.3 (62.3)	163.5 (55.3)	151.2 (57.9)
Other CGM	136.0 (53.2)	142.5 (52.7)	147.2 (54.0)	195.4 (65.7)	142.7 (46.0)	155.1 (66.6)	142.8 (48.3)	192.8 (56.7)	142.0 (56.9)	164.2 (47.7)	158.1 (58.2)
KS Test $p < 0.05$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

[a] Abbreviations: CGM, Continous Glucose Monitoring; KS, Kolmogorov–Smirnov.

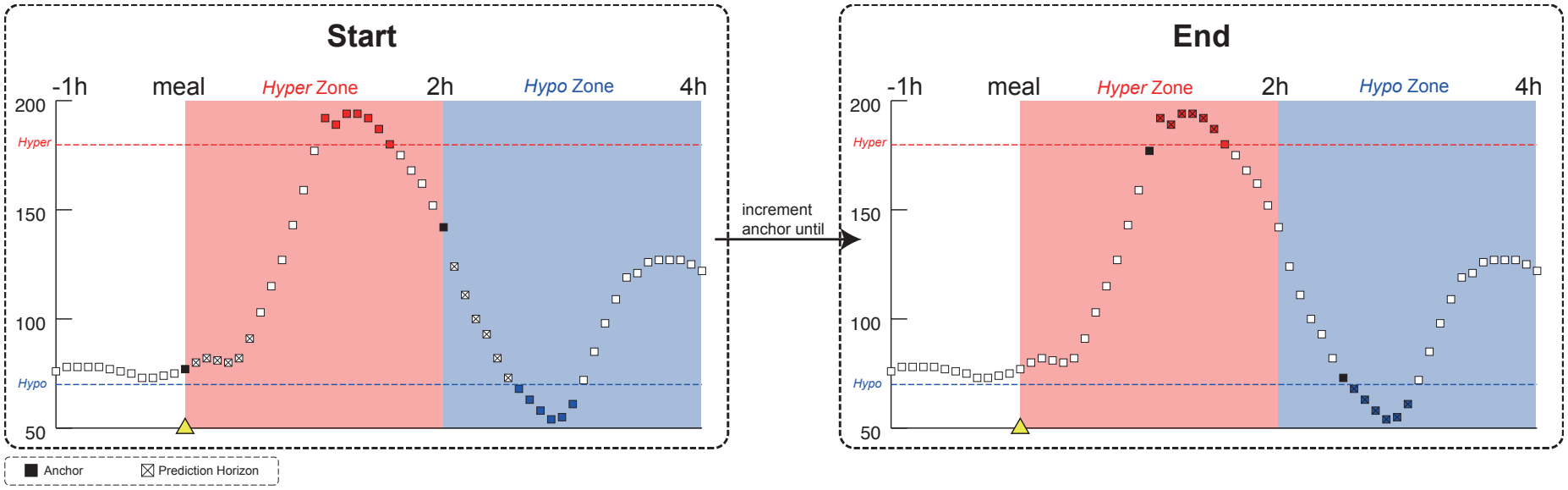


Figure 6.2: The illustration of the training and testing examples extraction. Taking hyperglycemia for illustration, an anchor started to increment from the first timestamp in the red zone, and a corresponding example was generated at each incrementing step until it arrived just before the occurrence of a hyperglycemia event. The extraction of hypoglycemia was of a similar manner as illustrated in the blue zone. Figure is adapted from Cui, Nolan et al. (2023).

6.2.1 Problem Definition

Under the modern CGM configuration with a 5-minute sampling rate, the prediction horizon was set to be 30 minutes (corresponding to 6 data points in the experiments of this study). In accordance with clinical consensus (Daenen et al., 2010; Altuntaş, 2019), 0-2 hours and 2-4 hours after the meal were considered to be the effective time range for the two tasks of postprandial hyperglycemia and hypoglycemia, respectively. Under this configuration, the 4-hour postprandial glucose trajectory, denoted by $\mathbf{g} = [g_1, g_2, ..., g_{48}]$, was the source of examples extraction for a meal at the time of sample g_0. Specifically, for each task, I applied a moving anchor i starting at the first index of its time range. A new training/testing example (i, y_i) was generated each time the anchor incremented until the prediction horizon moved to the last 6 data points in the time range. To this end, the anchor i took a default range in

$$\text{hyperglycemia} \qquad \{i \in \mathbb{N} : 1 \leq i \leq 18\}, \tag{6.1}$$

$$\text{hypoglycemia} \qquad \{i \in \mathbb{N} : 25 \leq i \leq 42\}. \tag{6.2}$$

It was defined that a hyperglycemia (hypoglycemia) event occurred if at least two consecutive data points in the prediction horizon crossed the threshold of 180 mg/dL (70 mg/dL) (Bergenstal et al., 2013). Consequently, the label y_i was a binary value determined by the occurrence of the event within the prediction horizon $\mathbf{g}_{i+1:i+6}$.

On this basis, an early stopping criterion in the incremental process of the anchor i was applied: taking the generation of the hyperglycemia instances as example, if the hyperglycemia event did not occur in the effective time range, i.e., $[g_1, ..., g_{24}]$, then no early stopping was applied and all the 18 instances were extracted; if the hyperglycemia occurred in $[g_1, ..., g_{24}]$, then the anchor i early stopped at the last index before the occurrence of that event. This process is illustrated in Figure 6.2. This was to reflect that the prediction corresponding to a certain meal was no longer required once the meal already caused hyperglycemia (hypoglycemia). By repeating the procedure described above on all meals, the full postprandial dataset could be constructed using the raw personalised CGM data. A pseudo code snippet for this training and testing examples generation is shown in Algorithm 1.

6.2.2 Classification to Regression

Under this problem formulation, a central component of the proposed methodology was to change the natural binary classification setup adopted by previous studies (Oviedo, Contreras et al., 2019; Seo, Lee et al., 2019; Vehí et al., 2020) into an approximated regression scheme. More specifically, instead of directly making a prediction of the binary y_i, I let the model learn to predict a numerical value of the max (min) glucose value in the prediction horizon, then the binary decision was made via the threshold 180 mg/dL (70 mg/dL). Note that this was only a methodological innovation, meaning this newly defined supervision signal was only used in the training phase and model inference, while in the evaluation phase the ground truth remained using the binary label defined in Section 6.2.1. Although this translation was not exactly equivalent to the definition of two-consecutive points crossing the threshold, this scheme was employed for its stronger supervision to the model training, because the binary supervision would lose the information of the severity of the hyperglycemia (hypoglycemia) event. Moreover, as the abnormal-glycemia events tended to be rare or even did not present at all (especially for hypoglycemia), the positive and negative examples tended to be significantly imbalanced. This regression scheme also naturally avoided the need of dealing with the imbalanced class problem, such as choosing a proper class weighting in classification loss.

Algorithm 1 Pseudo Codes for Extracting Train/Test Examples from a Meal.

```
1: Input: g, t                                          ▷ CGM sequence, meal time
2: Initialise ph = 6                                    ▷ prediction horizon: 0.5 h
3:
4: Case 1: hyperglycemia
5: Initialise thres = 180
6: Initialise T = [1, 2, ..., 18]                       ▷ 0-2 h
7: procedure ExtractData(g, t)
8:     data = []
9:     for i in T do
10:        if g_{t+i} ≥ thres then
11:            break
12:        end if
13:        if two consecutive in g_{(t+i+1):(t+i+ph)} ≥ thres then
14:            y = True
15:        else
16:            y = False
17:        end if
18:        push tuple (t, i, y) to data
19:    end for
20:    return data
21: end procedure
22:
23: Case 2: hypoglycemia
24: Initialise thres = 70
25: Initialise T = [25, 26, ..., 42]                    ▷ 2-4 h
26: procedure ExtractData(g, t)
27:    Change ≥ to ≤ in line 10 and 13.
28: end procedure
```

6.2.3 Unified Model and Joint Training for Hyperglycemia and Hypoglycemia

As the definition of hyperglycemia and hypoglycemia only differed in the threshold and time range, I approached the two tasks simultaneously using a unified model structure which took the recent glucose pattern as input. To be specific, I used the 1-hour recent CGM sequence $\mathbf{g}_{i-11:i}$ as the input and left out all other factors such as the meal and insulin amount in consideration of the difficulty of obtaining reliable meal amount estimation (Brew-Sam et al., 2021) and that the data may not be sufficiently large to capture the personalised effect of insulin. To encode the sequential feature, I fed the input 1-hour CGM sequence into an LSTM (Hochreiter and Schmidhuber, 1997) backbone with hidden dimension D followed by a ReLU activation function (Nair and Hinton, 2010). The last hidden state $\mathbf{h}$ of the LSTM, given by

$$\mathbf{h} = \max(\text{LSTM}(\mathbf{g}_{i-11:i}), 0) \in \mathbb{R}^D, \tag{6.3}$$

was regarded as the encoded feature vector of the input CGM sequence. The LSTM backbone was followed by two linear heads, which took the identical $\mathbf{h}$ as the input, but were respectively responsible for the prediction of hyperglycemia and hypoglycemia. To reflect the classification to regression scheme, the two linear heads projected $\mathbf{h}$ to two numerical outputs that respectively corresponded to the maximum and minimum glucose in the prediction horizon. The two predictions were given by

$$\hat{g}_{\text{hyper}} = \mathbf{W}_{\text{hyper}} \cdot \mathbf{h} + b_{\text{hyper}} \in \mathbb{R}, \tag{6.4}$$

$$\hat{g}_{\text{hypo}} = \mathbf{W}_{\text{hypo}} \cdot \mathbf{h} + b_{\text{hypo}} \in \mathbb{R}, \tag{6.5}$$

where $\mathbf{W} \in \mathbb{R}^D$ and $b \in \mathbb{R}$ represented the canonical linear weight and bias term in a linear projection, respectively.

Finally, mean squared error (MSE) loss between the predicted value and the ground truth was employed in supervising the model training, given by

$$\mathcal{L}_{\text{hyper}} = \frac{1}{M}\sum^{M}(\hat{g}_{\text{hyper}} - \max \mathbf{g}_{i+1:i+6})^2, \tag{6.6}$$

$$\mathcal{L}_{\text{hypo}} = \frac{1}{N}\sum^{N}(\hat{g}_{\text{hypo}} - \min \mathbf{g}_{i+1:i+6})^2, \tag{6.7}$$

where M and N denoted the total number of training examples extracted for hyperglycemia and hypoglycemia, respectively.

As the two linear heads shared the same hidden state $\mathbf{h}$ as the input, the LSTM was encouraged to learn the temporal trend of the recent glucose without bias towards either task, such as predicting a high glucose value for hyperglycemia or a low glucose value for hypoglycemia. To promote this in the training process, in each training iteration, a hyperglycemia batch and a hypoglycemia batch were independently fed to the model, and the model was alternatively updated using the corresponding loss (i.e., $\mathcal{L}_{\text{hyper}}$ or $\mathcal{L}_{\text{hypo}}$, depending on the task of that batch). As will be shown in the ablation study in Section 6.3, the jointly designed methodology for the two tasks hyperglycemia and hypoglycemia achieved comparable performance to the default option of independently training two models for the two tasks, yet half of the computational resources were saved.

6.3 Experiments

In this section, the evaluation method is presented in Section 6.3.1. The results of the experiments, including three ablation studies (described in Section 6.3.2), are presented in Section 6.3.3.[1]

6.3.1 Evaluation Method

The proposed methodology was evaluated using the OhioT1DM dataset under a personalised training setup, i.e., without instance adaptation as in Chapter 4 and 5. The evaluation metrics used in this study included the following: 1) the RMSE was used for evaluating the regression output in accordance with existing studies in general glucose prediction, 2) several metrics calculated out of the confusion matrix for the final classification, namely the Sensitivity (SE), Specificity (SP), and False Alarm Rate (FA) were further evaluated, and 3) MCC was employed as an overall reflection of the classification result. Additionally, to evaluate the timeliness of the prediction, the Detection Time (DT) of the start of a hyperglycemia or hypoglycemia event averaged over the meals that were correctly predicted by the algorithm was calculated. In these metrics, MCC took values in $[-1, 1]$ to reflect the overall classification quality and others in $[0, 1]$, larger (smaller) values reflected better performance for SE, SP, and MCC (FA). For DT, the theoretically optimal value was 30 minutes as it was the preset prediction horizon. To enable statistical significance testing, I repeated all experiments 10 times with different random seeds which would affect the intialisation of the model trainable parameters, and reported the averaged scores and their standard deviation in all tables.

[1] Dataset details can be found in Appendix A. Implementation details can be found in Appendix B.

Table 6.2: Main performance comparison in the format of mean (std) over 10 repeated experiments with different random seeds. Bold scores were statistically significantly better ($p < 0.05$) than all baselines via the unpaired t test.

	Hyperglycemia						Hypoglycemia					
	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑
All-day (Deng, Lu et al., 2021)	19.08	-	-	-	-	-	19.08	-	-	-	-	-
All-day (Yang, Yu et al., 2022)	18.93	-	-	-	-	-	18.93	-	-	-	-	-
All-day (Cui, Hettiarachchi et al., 2021)	17.97	-	-	-	-	-	17.97	-	-	-	-	-
Postprandial (Seo, Lee et al., 2019)	-	-	-	-	-	-	-	0.90	0.91	0.70	-	25.5
								(0.03)	(0.02)	(0.04)		(1.97)
Postprandial (Vehí et al., 2020)	-	-	-	-	-	-	-	0.69	0.8	-	0.24	-
Postprandial (Oviedo, Contreras et al., 2019)	-	-	-	-	-	-	-	0.71	0.79	-	0.48	-
Replication (Seo, Lee et al., 2019)	-	0.83	0.81	0.58	0.50	19.28	-	0.82	0.94	0.85	0.34	19.4
		(0.01)	(0.00)	(0.01)	(0.00)	(0.23)		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Baseline - Dummy	27.01	0.03	1.00	0.00	0.15	5.00	17.44	0.05	1.00	0.40	0.17	5.00
Baseline - AdaBoost	22.03	0.71	0.88	0.46	0.53	16.53	17.42	0.07	1.00	0.70	0.13	10.25
	(0.07)	(0.01)	(0.01)	(0.01)	(0.01)	(0.28)	(0.14)	(0.02)	(0.00)	(0.09)	(0.03)	(2.36)
Baseline - Random Forest	19.45	0.58	0.94	0.33	0.55	14.65	14.14	0.31	1.00	0.48	0.39	8.93
	(0.04)	(0.01)	(0.00)	(0.01)	(0.01)	(0.19)	(0.02)	(0.01)	(0.00)	(0.01)	(0.01)	(0.23)
Baseline - MLP	18.79	0.57	0.94	0.33	0.55	14.79	13.23	0.38	0.99	0.51	0.42	8.94
	(0.21)	(0.01)	(0.00)	(0.01)	(0.01)	(0.28)	(0.19)	(0.05)	(0.00)	(0.04)	(0.02)	(1.54)
Proposed approach	**18.23**	0.66	0.95	0.33	**0.61**	15.01	13.25	0.48	0.99	0.50	**0.48**	10.97
	(0.35)	(0.04)	(0.01)	(0.03)	(0.02)	(0.98)	(0.17)	(0.05)	(0.01)	(0.05)	(0.02)	(0.89)

[a] Abbreviations: RMSE, Root Mean Squared Error (mg/dL); SE, Sensitivity; SP, Specificity; FA, False Alarm; MCC, Matthews's Correlation Coefficient; DT, Detection Time; MLP, Multi-layer Perceptron.

6.3.2 Ablation Studies

To further evaluate the validity of the proposed method, I conducted three ablation studies targeting on different arguments in the method.

Individual Training This ablation experiment aimed at investigating the validity of the joint model for hyperglycemia and hypoglycemia against treating the two tasks separately. Thus, instead of training the model alternatively using data from the two tasks as in the proposed method, two individual models were trained and evaluated in this ablation study.

Binary Supervision Different from previous studies (Oviedo, Contreras et al., 2019; Seo, Lee et al., 2019; Vehí et al., 2020), the proposed method approached the classification output in the way of translating the predicted max or min glucose in the prediction horizon to binary result using the defined thresholds. To investigate its effectiveness, I conducted the ablation study of applying binary supervision. More specifically, the output module of the model was changed to a binary classification layer in this ablation experiment. Correspondingly, the training was supervised by cross-entropy loss with class weights estimated via the class distribution in training data to mitigate the effect of class imbalance, instead of using the MSE loss as in the proposed method.

All-day Training In this study, the postprandial glucose prediction was treated in separate to general glucose prediction due to the distribution shift from postprandial glucose to overall glucose as shown in Figure 6.1 and Table 6.1. To further investigate the effectiveness of this decision, I conducted the ablation experiment of training the model using examples extracted

Table 6.3: Ablation experiments in the format of mean(std) over 10 repeated experiments with different random seeds.

	Hyperglycemia						Hypoglycemia					
	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑
Individual Training	18.64	0.64	0.96	0.29	0.62	14.22	13.50	0.40	1.00	0.47	0.45	10.13
	(0.14)	(0.01)	(0.01)	(0.02)	(0.01)	(0.28)	(0.08)	(0.05)	(0.00)	(0.05)	(0.03)	(0.79)
Binary Supervision	-	0.80	0.80	0.60	0.47	18.11	-	0.70	0.90	0.90	0.24	15.7
		(0.02)	(0.01)	(0.01)	(0.01)	(0.54)		(0.06)	(0.05)	(0.04)	(0.06)	(0.80)
All-day Training	19.67	0.51	0.97	0.25	0.57	11.71	12.88	0.30	1.00	0.52	0.37	7.13
	(0.21)	(0.02)	(0.01)	(0.02)	(0.01)	(0.38)	(0.15)	(0.04)	(0.01)	(0.09)	(0.03)	(0.98)
Proposed approach	18.23	0.66	0.95	0.33	0.61	15.01	13.25	0.48	0.99	0.50	0.48	10.97
	(0.35)	(0.04)	(0.01)	(0.03)	(0.02)	(0.98)	(0.17)	(0.05)	(0.01)	(0.05)	(0.02)	(0.89)

[a] Abbreviations: RMSE, Root Mean Squared Error (mg/dL); SE, Sensitivity; SP, Specificity; FA, False Alarm; MCC, Matthews's Correlation Coefficient; DT, Detection Time.

from full CGM data which was consistent to the setting of short-term glucose prediction, yet the testing data were kept unchanged as only postprandial.

6.3.3 Results

As shown in Table 6.2,the proposed method was compared to state-of-the-art studies in the *all-day glucose prediction* field and the *postprandial glucose prediction* field. Moreover, the main body of the comparison was several baselines using traditional ML algorithms.

Comparing to All-day Glucose Prediction Studies As shown in the first section of Table 6.2, the studies Cui, Hettiarachchi et al. (2021); Deng, Lu et al. (2021); Yang, Yu et al. (2022) were all conducted using the same OhioT1DM dataset while evaluated using RMSE between the actual and predicted glucose value without including the other metrics used in this study as the prediction of abnormal-glycemia events was out of their scope. For RMSE, this study was slightly different to those studies as the regression target was the max and min glucose in the next 30 minutes instead of the exact glucose at 30 minutes ahead. However, the proposed methodology achieved a comparable RMSE to these studies when predicting hyperglycemia and an RMSE significantly better than these studies when predicting hypoglycemia, suggesting a performance no lower than the methodologies proposed in these studies if applied to this task.

Comparing to Postprandial Glucose Prediction Studies Among the postprandial glucose prediction studies in the performance comparison in the second section of Table 6.2, Seo, Lee et al. (2019) followed the fixed prediction horizon scheme with a 30-minute prediction horizon, while Oviedo, Contreras et al. (2019); Vehí et al. (2020) adopted the meal-time prediction scheme. These studies were conducted on datasets unavailable to the community due to restrictions of the clinical trials, which made our comparison less straightforward. However, though a lower DT was presented, the proposed method achieved the state-of-the-art MCC, indicating an promising overall performance of the prediction.

Comparing to Baselines To enhance the comparative assessment, I replicated the methodology proposed in Seo, Lee et al. (2019) and implemented several classical ML algorithms as shown in the third section of Table 6.2. All results here were based on the same dataset and

experimental conditions. To be specific, I first implemented a dummy model, which simply treated the most recent known glucose value as the regression output $\hat{g}_{\text{hyper}}$ and $\hat{g}_{\text{hypo}}$. As the dummy model had no ability to learn anything, no 10-fold experiment was conducted. This model was used to set up the theoretical lower bound of performance and to show that other methods had the learning ability to an extent. On this basis, I implemented several traditional regression algorithms, namely random forest (RF), AdaBoost, and multi-layer perceptron (MLP) using the same 1-hour past CGM sequence as input, and further replicated the methodology proposed by Seo, Lee et al. (2019) which was an RF model using explicit features calculated out of the past CGM. In comparison with these baseline strategies, the proposed methodology achieved the highest MCC and lowest RMSE, and comparable performance on other metrics.

The results for the ablation studies are presented in Table 6.3. In the first ablation study (the individual training), my unified model achieved results close to the individual training scheme in all metrics. This observation suggested the capability of the shared LSTM backbone to properly encode the temporal feature of the input glucose history which was then used for prediction on the two tasks. By unifying the two tasks into one model, the need for computational resources was minimised both in inference time and memory requirement. In the binary supervision ablation study, though the DT was increased, this training strategy led to almost double FA and caused a significant drop of overall predictive performance ($\sim$12% MCC), indicating that the numerical supervision is a more effective strategy. In the ablation study that changed the training to all-day data, this change not only did not help increase the overall performance, but rather caused a deterioration in both the MCC and DT. Additionally, the full training data were approximately of 10 times larger amount than postprandial training data, meaning that invoking a large amount of less relevant training data had only misled the learning. These observations indicated the validity of the proposed problem definition for postprandial glucose prediction under the fixed horizon scheme and the effectiveness of the proposed method.

6.4 Discussion

The contributions of this study are summarised as follows:

- I proposed for the first time to formulate the problem of postprandial glucose prediction to be a joint problem of predicting postprandial hyperglycemia and hypoglycemia. The proposed ML method that unified the two targets into one single model achieved state-of-the-art prediction quality compared to existing studies, and also minimised the need for computational resources.
- The statistical analysis and ablation study verified the impact of the glucose distribution shift problem in postprandial glucose prediction, thus necessitated the research of treating the problem differently than the all-day glucose prediction.

This study identified the distribution shift of glucose data in the postprandial scenario, and focused on the problem of postprandial hyperglycemia and hypoglycemia prediction. For the first time I formulated the problem of joint prediction of postprandial hyperglycemia and hypoglycemia, and proposed a method using a simple LSTM backbone as a pioneering solution. The limitation of this study is the exclusion of meal-related factors such as carbohydrate amount and insulin dosing into the model. Such exogenous input could have more influence to the predictive performance than a regular all-day glucose prediction system, as the dealt problem focuses on the postprandial scenario. Hence, this is considered to be a vital direction for future work.

This study targeted the "engineering" objective of this study from the perspectives of training & testing examples extraction and model efficiency. Next, I will evaluate additional aspects of the engineering details in a ML-based glucose prediction pipeline to gain further insights in this field. This is to be discussed in Chapter 7.

Chapter 7

Evaluating the Engineering Options

This chapter is under review and planned to be submitted as

R. Cui, E. Daskalaki, C. J. Nolan and H. Suominen (2024). 'Evaluation of Different Engineering Options for Data-driven Glucose Prediction'. In: *TBD*

The corresponding work is open-sourced at `https://github.com/r-cui/EngineeringGluPred`.

Author contributions: R.C., E.D., C.J.N, and H.S. conceptualised the study; R.C. designed and performed research; R.C. wrote the manuscript and revised it, based on critical comments by E.D., C.J.N., and H.S.

This study related to the "engineering" objective of the thesis, as introduced in Section 1.2.3. Research in the field of data-driven glucose prediction is focusing heavily on the model design (Section 2.1), leading to that less attention is being paid to other aspects of the ML pipeline. Indeed, it is observed that different engineering settings have been adopted in different studies. For example, for the methodology, some studies have adopted the rolling prediction scheme (De Bois, Yacoubi et al., 2019; Cui, Hettiarachchi et al., 2021; Zaidi et al., 2021) and some have adopted direct prediction (Fox et al., 2018; Kim et al., 2020; Rabby et al., 2021; Koca et al., 2023); for the prediction target, existing studies differ in excursion prediction (Calm, Garcia-Jaramillo et al., 2007; Adelberger et al., 2021) or single-point prediction (Dong et al., 2019; Li, Liu et al., 2019; Mirshekarian et al., 2019). When implementing a ML pipeline for the purpose of data-driven glucose prediction, one needs to make an exclusive decision on such options. Motivated by these, I conducted a comparative evaluation of 1) rolling prediction vs. direct prediction, and 2) single-point prediction vs. excursion prediction, in order to analyse their pros and cons, and make suggestions on which one to adopt. With my experiments using vanilla models, where different engineering options have been changed as the sole modularised variable, it was concluded that the direct prediction methodology and the excursion prediction target are preferable than their counterparts.

7.1 Introduction

In the field of data-driven glucose prediction, as reviewed in Section 2.1, a dominant portion of existing studies focus on novel model architectures to increase the predictive performance. For example, Li, Daniels et al. (2019) proposed the convolutional recurrent neural network which focuses on a novel layer definition that combines the advantages of RNN and CNN. As shown in Figure 2.6, the focus of such studies lies in a central segment in a ML pipeline, that is the model segment. Although not gaining as much research attention, other components

in the ML pipeline can also influence the overall effectiveness. It is observed that existing studies differ in their implementation strategies with respect to various engineering details, and two of the most critical choices that need to be made are the prediction scheme and the prediction target. More specifically, the following two different options were considered in this study.

- Prediction scheme: rolling prediction vs. direct prediction.
- Prediction target: excursion prediction vs. single-point prediction.

As reviewed in Section 2.3.2, for the prediction scheme, some studies have adopted the rolling prediction scheme (De Bois, Yacoubi et al., 2019; Cui, Hettiarachchi et al., 2021; Zaidi et al., 2021) in which the full prediction horizon is reached by recursively predicting the next-step glucose and treating that value as a ground truth, while some have adopted direct prediction (Fox et al., 2018; Kim et al., 2020; Rabby et al., 2021; Koca et al., 2023) in which all required glucose values in the horizon are predicted all at once. The former coincides with the natural intuition of a step-by-step approach in glucose prediction, yet might suffer from the accumulated error problem (Hamzaçebi et al., 2009; Bengio, Vinyals et al., 2015) and the gradient vanishing/exploding problems (Pascanu et al., 2013) in training such a recursive model. The latter, on the other hand, provides a simple and straightforward way in doing the same task, but might be limited in terms of predictive performance. For the prediction target, some studies have performed excursion prediction (Calm, Garcia-Jaramillo et al., 2007; Adelberger et al., 2021) which requires predicting the glucose trajectory in a prediction horizon, while some have employed single-point prediction (Dong et al., 2019; Li, Liu et al., 2019; Mirshekarian et al., 2019) in which only the last-point glucose value in the horizon is predicted and evaluated. It is apparent that the former prediction target is a super set of the latter, thus being more informative and plausibly more challenging. Therefore, it raises a natural question of which option is more preferable. These options need to be considered in any data-driven glucose prediction strategy, however, there has not been a study focusing on comparing them. Besides, it is hard to draw clear conclusions on which option is better by simply cross comparing studies using different options, since inconsistency can exist in any component of their ML pipelines.

The objective of this study was to fill this gap by creating a pure and fair arena comparing the aforementioned different options that exist in the engineering of data-driven glucose prediction in terms of their predictive and efficiency performance. To investigate this, I conducted comparative experiments where the targeted engineering options were changed as the sole modularised variable which enabled the comparison. The experiment was repeated multiple times using the vanilla LSTM, GRU, and TCN models. Note that it was a deliberate choice to use these simple and widely investigated structures in this study, in order to minimise the potential unknown effects of complex model architectures. The experiments using the OhioT1DM dataset showed that under a modern setting of data-driven glucose prediction, direct prediction achieved better predictive performance and efficiency, and the two prediction targets exhibited no significant difference in the performance and efficiency. These results suggested that direct prediction might be a better option compared to rolling prediction, while excursion prediction would be beneficial as it provided more insight without loss of efficiency.

7.2 Methodology

This section first formally describes the definitions of the different engineering options for the prediction scheme in Section 7.2.1 and the prediction target in Section 7.2.2. The methodology

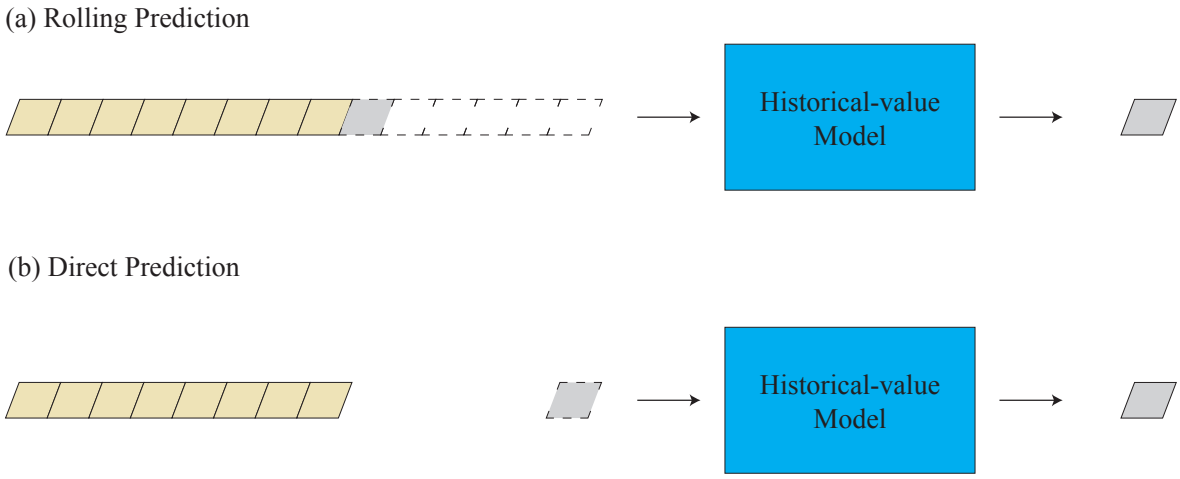


Figure 7.1: Comparing the two prediction schemes. (a) Illustrating the first step of the rolling prediction scheme, in which the predicted glucose value is appended to the end of the input sequence to enable the next prediction step. (b) Illustrating the direct prediction scheme, in which a desired glucose value is directly computed by a one-shot prediction. Note that only the single-point prediction target is presented in this diagram for illustration purpose, and the prediction scheme is independent to the prediction target, whether excursion or a single point.

evaluating and comparing these options is then narrated in Section 7.2.3.

7.2.1 Prediction Scheme

The prediction scheme difference appears in studies using historical-value models. The rolling prediction scheme (De Bois, Yacoubi et al., 2019; Cui, Hettiarachchi et al., 2021; Zaidi et al., 2021) produces the targeted glucose via recurrently predicting the next-step glucose using the past glucose sequence. That is, at time t, a rolling prediction learns to predict the next-step glucose via

$$\hat{g}_{t+1} = f([g_t, g_{t-1}, ...]).$$

The predicted value $\hat{g}_{t+1}$ is then treated as ground truth and appended to the end of the input sequence. Next, the new input sequence is recurrently fed to the model to predict the next value in time, until the prediction horizon is reached. More formally, the subsequent computation is given by

$$\begin{aligned}
\hat{g}_{t+2} &= f([\hat{g}_{t+1}, g_t, g_{t-1}, ...]),\\
\hat{g}_{t+3} &= f([\hat{g}_{t+2}, \hat{g}_{t+1}, g_t, g_{t-1}, ...]),\\
&\vdots\\
\hat{g}_{t+H} &= f([\hat{g}_{t+H-1}, ..., \hat{g}_{t+1}, g_t, g_{t-1}, ...]).
\end{aligned}$$

It can be seen from this process that the rolling prediction scheme intuitively models the natural elapse of time. However, it has been questioned for the accumulated error problem (Hamzaçebi et al., 2009; Bengio, Vinyals et al., 2015) i.e., the error made in each prediction step would be accumulated exponentially, and the gradient vanishing and exploding problems (Pascanu et al., 2013) which cause difficulty in training.

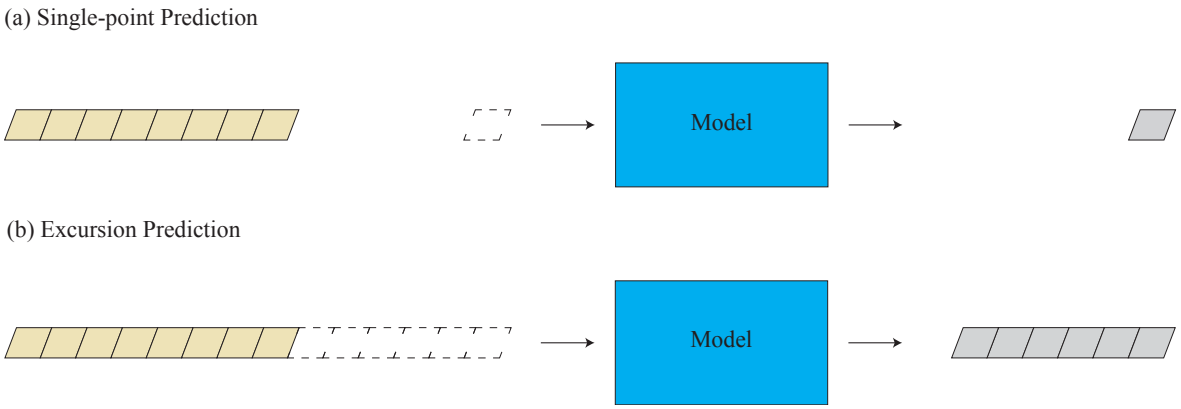


Figure 7.2: Comparing the two prediction targets. (a) In single-point prediction, the desired output is solely the last value in the prediction horizon. (b) In excursion prediction, the desired output is the full trajectory through the prediction horizon. Hence, the excursion target is a super set of single-point predictions.

In contrast, the direct prediction scheme, as used in studies like Fox et al. (2018); Kim et al. (2020); Rabby et al. (2021); Koca et al. (2023), learns a one-shot mapping via

$$\hat{g}_{t+H} = f([g_t, g_{t-1}, ...]).$$

Although not following the natural sampling rate in making the prediction, the direct prediction scheme holds the advantage of being much simpler than the rolling prediction scheme. An illustration of the difference between the two prediction schemes is shown in Figure 7.1. As applying the rolling scheme or the direct scheme is an exclusive choice in implementing a ML based glucose prediction pipeline, it is worth evaluating and comparing them to draw conclusion on which one is preferable.

7.2.2 Prediction Target

For the prediction target (Figure 7.2), a dominant number of existing studies applied the single-point prediction, which requires predicting the sole value

$$\hat{g}_{t+H}.$$

For example, Mirshekarian et al. (2019) have evaluated their LSTM-based neural attention models in predicting short-term glucose at the 30th minute and 60th minute. Similarly, Li, Liu et al. (2019) propose GluNet inspired by TCN and evaluates the predictive performance at the 30th minute and 60th minute. Dong et al. (2019) have reported the performance of their Clustered-RNN at the 30th, 45th, and 60th minute. A natural consequence of using this single-point prediction is that studies can only evaluate their results using numeric metrics such as RMSE, and whether the value falls into hyperglycemia ($\geq$ 180 mg/dL) or hypoglycemia ($\leq$ 70 mg/dL). The homogeneity of the evaluation metric has restricted the comprehensiveness of existing studies to make cross comparison with other studies. This limitation would be largely alleviated if the prediction target is the excursion, as in some studies like Calm, Garcia-Jaramillo et al. (2007); Adelberger et al. (2021), which requires predicting the full trajectory of the prediction horizon

$$[\hat{g}_{t+H}, ..., \hat{g}_{t+2}, \hat{g}_{t+1}].$$

Under excursion prediction, as was studied in Chapter 5, the performance can be evaluated via the numerical accuracy, the trend polarity, the accuracy of hyperglycemia and hypoglycemia prediction, and other potential metrics yet to be proposed. As an extended research target from single-point prediction, the excursion prediction would be more challenging, whereas I hypothesised that from the engineering perspective, its implementation would result in only a trivial increase in cost. Therefore, modern research should target on investigating the task of glucose excursion prediction, as a reliable excursion prediction would be more informative.

7.2.3 Ablation Setup

Ablation study (Meyes et al., 2019) was employed as the methodology of evaluating and comparing the different engineering options. That is, all experiments were based on a same ML pipeline, in which different engineering options were substituted into the pipeline as a modular component, and the corresponding performance was compared. To best ensure the purity of the experimental environment, the pipeline was designed to be standard and simple, i.e., apart from the engineering options, all the other components in the pipeline adopted the most reliable and simple setups in existing studies. More specifically, in all experiments, the dataset used was the 12 patients in the OhioT1DM dataset (Marling and Bunescu, 2020) which underwent the same preprocessing of linear interpolation; the task setup was set to a 1-hour prediction horizon using a 1-hour glucose history as the input; the combination of Adam optimiser (Kingma and Ba, 2014) and cosine annealing scheduler (Loshchilov and Hutter, 2016) was used in training the model, with the same settings of batch size and early stopping scheme; the computational device used was the same NVIDIA GeForce RTX 3090 GPU. Each single experiment was repeated three times using three reliable and robust backbone models, namely the LSTM (Hochreiter and Schmidhuber, 1997), the GRU (Chung et al., 2014), and the TCN (Bai et al., 2018). The model specifications were borrowed from existing literature (Fox et al., 2018; Zaidi et al., 2021). More specifically, the LSTM and GRU contained two layers with a hidden dimension of 512, the TCN contained 3 layers with a hidden dimension of 64. The two research targets, prediction scheme and the prediction target, were individually investigated following the described methodology. For simplicity, in the prediction scheme experiment, the single-point prediction target was adopted, and in the prediction target experiment, I adopted the direct prediction scheme. In the evaluation, the predictive performance was reported using the most widely adopted RMSE metric, and unpaired t test was used for statistical significance evaluation. The efficiency performance was reported using the model inference time, number of MAC operations, and the occupied GPU memory.

7.3 Experiments

This section presents the details of the results regarding the two research targets introduced in Section 7.3.1 and 7.3.2, respectively.

7.3.1 Prediction Scheme

This section discusses the results of comparing the predictive performance between the two prediction schemes. As shown in Table 7.1, by cross comparing the predictive performance of the same model under different prediction schemes, all models experienced a performance improvement from 2 to 3 mg/dL on all the 12 patients by changing the prediction scheme from rolling to direct, with the TCN experiment achieving statistical significance. This observation suggested that the accumulated error produced by the recursive prediction

Table 7.1: Comparing the two prediction schemes, rolling prediction vs. direct prediction, in measuring the RMSE (mg/dL) of predicting the 60-minute glucose using the OhioT1DM dataset. An underlined score indicates a statistical significant result ($p < 0.05$) in comparison with its counterpart under t test.

	Model	540	544	552	559	563	567	570	575	584	588	591	596	avg (std)
Rolling	LSTM	41.06	34.14	31.17	35.29	31.32	39.58	30.02	36.95	38.19	33.04	35.4	30.29	34.70 (3.54)
	GRU	42.45	33.68	32.57	36.57	31.62	40.72	30.7	37.67	37.46	33.51	35.92	30.38	35.27 (3.71)
	TCN	43.1	34.37	31.81	37.44	31.42	41.82	31.42	37.71	40.76	34.12	36.11	32.5	36.05 (3.98)
Direct	LSTM	39.95	31.3	31.16	34.71	30.94	38.04	28.29	36.56	36.37	31.56	33.96	29.4	33.52 (3.49)
	GRU	39.66	31.44	30.75	34.66	31.24	37.35	28.24	37.64	36.88	32.44	33.68	29.43	33.62 (3.47)
	TCN	39.48	31.11	30.01	33.79	30.02	37.17	27.44	35.17	35.99	32.03	33.34	29.13	32.89 (3.44)

[a] Abbreviations: LSTM, Long Short-term Memory; GRU, Gated Recurrent Unit; TCN, Temporal Convolutional Network.

Table 7.2: Comparing the two prediction schemes, rolling prediction vs. direct prediction, in efficiency. The inference time is the average of 1000 runs on different mini-batches.

	Model	Inference Time	MACs	GPU Memory
Rolling	LSTM	2.04ms	4.658G	1473MiB
	GRU	2.47ms	3.495G	1437MiB
	TCN	15.06ms	635.019M	1811MiB
Direct	LSTM	0.42ms	2.430G	1429MiB
	GRU	0.41ms	1.823G	1409MiB
	TCN	0.99ms	38.433M	1799MiB

[a] Abbreviations: LSTM, Long Short-term Memory; GRU, Gated Recurrent Unit; TCN, Temporal Convolutional Network; MAC, Multiply-accumulate Operation.

scheme may be a potential drawback in data-driven glucose prediction. Not only with higher predictive performance, the direct prediction scheme also achieved better efficiency as shown in Table 7.2. More specifically, direct prediction scheme contributed to a 5 to more than 10 times reduction in model inference time. With this scheme, the feedforward propagation of the models was much simpler, reflected by the significantly lower number of MAC operations. To this end, the runtime occupied GPU memory of model in direct prediction scheme was also lower. These results suggest that under the modern problem setup of data-driven glucose prediction, a direct prediction scheme can be more preferable than the rolling prediction scheme in both predictive performance and efficiency.

7.3.2 Prediction Target

This section discusses the results of comparing the predictive performance between the two prediction targets. As shown in Table 7.3, where only the last point performance was evaluated, as that was the only mutual part between the two experiment groups, the two prediction targets were demonstrated to exhibit no significant difference in predictive performance. This was as expected since in practice, the predictions at different positions in the excursion are normally individually produced using different linear heads. The

Table 7.3: Comparing the two prediction targets, single-point prediction vs. excursion prediction, in measuring the RMSE (mg/dL) of predicting the 60-minute glucose using the OhioT1DM dataset.

	Model	540	544	552	559	563	567	570	575	584	588	591	596	avg (std)
Single	LSTM	39.95	31.3	31.16	34.71	30.94	38.04	28.29	36.56	36.37	31.56	33.96	29.4	33.52 (3.49)
	GRU	39.66	31.44	30.75	34.66	31.24	37.35	28.24	37.64	36.88	32.44	33.68	29.43	33.62 (3.47)
	TCN	39.58	31.21	30.1	33.89	30	37.27	27.54	35.27	36.09	32.13	33.54	29.23	32.99 (3.45)
Excursion	LSTM	40.4	31.87	29.92	35.29	30.52	36.25	29.13	38.14	36.84	31.82	34.03	29.74	33.66 (3.55)
	GRU	41.34	31.66	30.47	34.58	31.79	37.27	27.76	36.65	37.47	32.64	34.56	29.34	33.80 (3.75)
	TCN	39.56	31.55	30.05	33.93	29.86	37.14	27.42	35.9	36.57	31.69	33.97	29.1	33.06 (3.56)

[a] Abbreviations: LSTM, Long Short-term Memory; GRU, Gated Recurrent Unit; TCN, Temporal Convolutional Network.

Table 7.4: Comparing the two prediction targets, single-point prediction vs. excursion prediction, in efficiency. The inference time is the average of 1000 runs on different mini-batches.

	Model	Inference Time	MACs	GPU Memory
Single	LSTM	0.42ms	2.430G	1429MiB
	GRU	0.41ms	1.823G	1409MiB
	TCN	0.99ms	38.433M	1799MiB
Excursion	LSTM	0.43ms	2.430G	1431MiB
	GRU	0.41ms	1.824G	1409MiB
	TCN	0.99ms	38.478M	1799MiB

[a] Abbreviations: LSTM, Long Short-term Memory; GRU, Gated Recurrent Unit; TCN, Temporal Convolutional Network; MAC, Multiply-accumulate Operation.

efficiency of the two prediction targets was also evaluated as shown in Table 7.4, and very similar efficiency was observed. In other words, trivial engineering cost was added when the prediction target changed from single-point to the full excursion. Therefore, it can be concluded that further research could take the excursion prediction as the default target instead of the widely adopted single-point prediction, since there existed no significant trade-off, while the former is more informative. Furthermore, a per-position predictive performance of solely the excursion prediction experiment is provided in Table 7.5 and Figure 7.3. A stable 3 mg/dL increase of the RMSE was observed as the prediction position moved, indicating a steady increasing difficulty under the 1 hour prediction horizon.

7.4 Discussion

This section discusses the principal findings, the limitation, and the broader impact in the following subsections.

7.4.1 Principal Findings

This study evaluated and compared the different approaches in two existing engineering aspects, the prediction scheme and the prediction target. It was concluded that under the

current setting of data-driven glucose prediction, the direct prediction might serve better as the prediction scheme than rolling prediction as the former achieved better predictive performance and efficiency; and the excursion prediction should be prioritised as the prediction target since it demanded no additional computational resources and did not affect the performance of single-point predictions, while offering greater value and relevance to stakeholders.

7.4.2 Limitation

Apart from the two engineering options in this chapter, there exists a range of other options that need to be set in a glucose prediction problem formulation such as the use of exogenous inputs like insulin and food intake information. However, as researchers have proposed many different ways of using these inputs ranging from physiological models and data-driven physiological models, conducting a comparative experiment for the exogenous inputs is not as straightforward and intuitive as the discussed prediction scheme and prediction target which both involve only a simple twofold comparison. Hence, it was not included in the scope of this study and could be better investigated individually.

7.4.3 Significance and Broader Impact

With the specific concentration on engineering details for data-driven glucose prediction, this study aimed to assist ML practitioners making better decision in engineering their glucose prediction pipelines and avoiding unnecessary pitfalls that could affect the overall performance in the implementation. Additionally, it seeked to support ongoing research by targeting the more challenging goal of glucose excursion prediction, which could be of greater benefit to stakeholders. I hope the codes release openly accessible to public could facilitate broader advancements in related fields.

The contributions of this study are summarised as follows:

- It was demonstrated that under the modern setting of ML based data-driven glucose prediction, a direct prediction achieved higher predictive performance and lower resource consumption, thus being more preferable in the engineering.
- It was demonstrated that setting the prediction target to the full excursion in the prediction horizon would not affect the predictive performance of the single-point prediction, and required trivial extra cost, thus should be used in further research as it is more informative and challenging.

In this study, I focused on the practical engineering aspects of data-driven glucose prediction and evaluated two topics which existing studies have divergence of choices. For the prediction scheme, it was shown that direct prediction appeared more desirable as it achieved both higher predictive performance and efficiency than rolling prediction. For the prediction target, it was shown that research should take the excursion prediction as the target as it offered significantly more information than single-point prediction with no added complexity.

This chapter closes all the presentation of my original studies in this thesis. Next, Chapter 8 will conclude the thesis by summarising the principal findings, analysing the limitations, and proposing potential future works.

Table 7.5: The per-position RMSE for excursion prediction using the OhioT1DM dataset. Scores are given by averaging the performance on all the 12 patients in the dataset.

Model	5	10	15	20	25	30	35	40	45	50	55	60
LSTM	5.67 (1.23)	8.56 (1.45)	11.51 (1.69)	14.45 (1.89)	17.28 (2.06)	19.92 (2.29)	22.55 (2.49)	25.09 (2.65)	27.46 (2.92)	29.67 (3.13)	31.77 (3.35)	33.66 (3.55)
GRU	5.78 (1.26)	8.60 (1.42)	11.48 (1.67)	14.47 (1.93)	17.34 (2.07)	20.05 (2.18)	22.68 (2.46)	25.16 (2.70)	27.61 (3.00)	29.78 (3.26)	31.85 (3.51)	33.80 (3.75)
TCN	4.77 (1.08)	8.02 (1.42)	11.15 (1.70)	14.13 (1.88)	16.96 (2.05)	19.66 (2.25)	22.27 (2.52)	24.71 (2.70)	26.99 (2.87)	29.18 (3.15)	31.14 (3.32)	33.06 (3.56)

[a] Abbreviations: LSTM, Long Short-term Memory; GRU, Gated Recurrent Unit; TCN, Temporal Convolutional Network; RMSE, Root Mean Square Error.

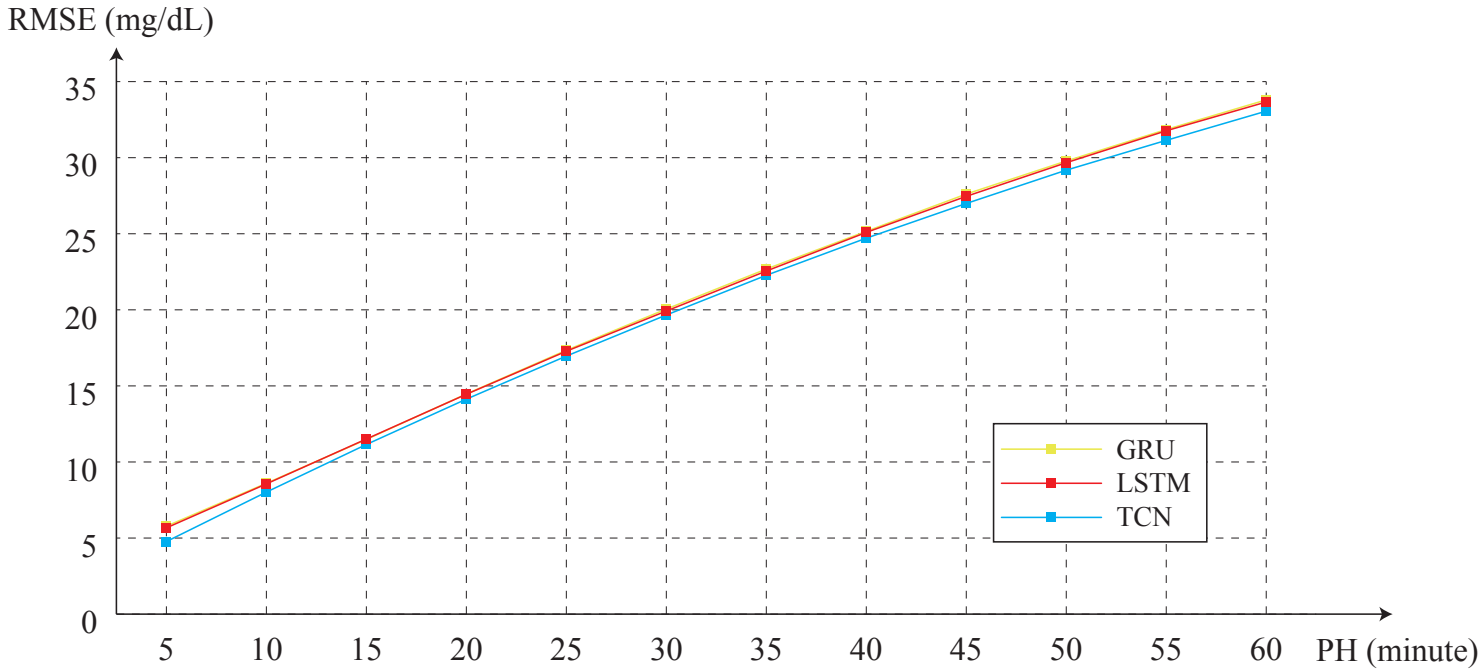


Figure 7.3: The per-position RMSE for excursion prediction using the OhioT1DM dataset. This is a visualisation of results shown in Table 7.5.

Chapter 8

Conclusion

This chapter concludes this thesis by summarising the principal findings, analysing the limitations and broader impacts, proposing future works, rethinking the broad field of data-driven glucose prediction, and giving final remarks.

8.1 Principal Findings

This thesis has aimed at improving data-driven glucose prediction for T1DM in the direction of modeling, adaptation, and engineering, and the five individual studies presented in the previous chapters are related to these topics in the accordance shown in Figure 1.2. Following same practice, this section summarises the principal findings of this thesis as follows.

1. **Modeling**
    - The development of the RSAN model demonstrated statistically significant improvement in predictive performance over traditional RNN and CNN-based models which are the state-of-the-art historical-value models.
    - The MOTIM demonstrated the potential of time-index-based modeling in glucose prediction. By reducing the complexity of the learning task, MOTIM offered comparable predictive performance to traditional methods with significantly improved efficiency.
2. **Adaptation**
    - The use of transfer learning with the RSAN model validated the effectiveness of instance adaptation using the RSAN model. Fine-tuning pre-trained models on individual patient data showed measurable improvements in prediction accuracy.
    - The application of meta-learning principles within the MOTIM framework provided a robust approach for instance adaptation, facilitating personalised glucose predictions with limited patient-specific data.
3. **Engineering**
    - The MOTIM model achieved significantly higher efficiency than existing RNN-based, CNN-based, and Transformer-based historical-value models by its core technique of time-index modeling, facilitating its usability in deployment to resource-limited devices.
    - A strategy of extracting the training & testing instances from daily CGM data for the purpose of postprandial glucose prediction was proposed. This could help standardise evaluation procedures and improve the accuracy of predictions.

- A detailed evaluation of different engineering options highlighted the benefits of direct prediction schemes and excursion prediction targets over their counterparts. These choices would not only enhance predictive accuracy but also improve the efficiency and practicality of data-driven glucose prediction implementations.

In summary, this research has successfully advanced the field of data-driven glucose prediction for T1DM by developing novel modeling techniques, implementing effective adaptation methods, and optimising engineering practices, thereby enhancing prediction accuracy, adaptability, and computational efficiency.

8.2 Comparison to Related Works

This section extends the principal findings of this thesis with a comparative analysis to existing research in data-driven glucose prediction. By analysing the similarities and differences, this comparison makes the contributions in this thesis more tangible and measurable.

The contribution of this thesis on modeling corresponds to the use of the RSAN and the MOTIM, described in Chapter 4 and 5. The proposal of the RSAN followed the existing paradigm of historical-value modeling, while it applied the novel self-attention mechanism as the core modeling for the first time. This model outperformed the state-of-the-art (Bertachi et al., 2018; Martinsson, Schliep, Eliasson, Meijner et al., 2018; Mirshekarian et al., 2019; Nemat et al., 2020; Zhu, Li, Chen et al., 2020; Zhu, Yao et al., 2020) at the time of the study, e.g., around 1mg/dL lower RMSE in predicting 30-minute glucose on the OhioT1DM dataset. The MOTIM method, on the other hand, shifted the direction of data-driven glucose prediction research towards time-index modeling, offering a novel approach distinct from the traditional historical-value models. As the first study in this new direction, the MOTIM achieved comparable performance to the existing studies (including the RSAN) (Fox et al., 2018; Zaidi et al., 2021) when using CGM records as the only source of input, e.g., 1-hour excursion averaged MAPE of 10.25 on OhioT1DM and 12.50 on UMT1DM. With these promising results, the feasibility of using these two new approaches to data-driven glucose prediction were validated, offering new opportunities for future research and practical applications in personalised glucose management.

For the adaptation of data-driven glucose prediction, related discussion of this thesis is also depicted in Chapter 4 and 5. More specifically, as a novel instance of historical-value models, the RSAN model was evaluated in the adaption scenario using the pre-training & fine-tuning method, which has been investigated by other historical-value based methods (Deng, Chen et al., 2020; De Bois, El Yacoubi et al., 2021; Deng, Lu et al., 2021; Yu, Yang et al., 2021). With the around 0.5 mg/dL and 1mg/dL RMSE improvements for the 30-minute and 60-minute prediction horizon, it was demonstrated that pre-training using data from other patients was helpful to the personalised training of an RSAN model. Moreover in Chapter 5, the MOTIM model further enhanced its synergy to the adaptation task by its inherent meta-optimisation framework. In the evaluation using the same experiment condition with existing methods, i.e., using the pre-training & fine-tuning scheme on MOTIM or historical-value models (Fox et al., 2018; Zaidi et al., 2021), the MOTIM achieved state-of-the-art performance (e.g., a 10.24 average MAPE on 60-minute prediction horizon on OhioT1DM) and the fastest training convergence (i.e., minimal performance gap from $PoU = 0$ to $PoU = 1$). Through these results, the proposal of the novel modeling to data-driven glucose prediction in this thesis have not only demonstrated substantial improvements in prediction accuracy but also shown effectiveness in adaptation tasks.

For the engineering objective, related discussion of this thesis is presented in Chapter 5, 6, and 7. The MOTIM model presented in Chapter 5 was demonstrated significantly more efficient than existing historical-value models due to its fundamental innovation of time-index based modeling. Comparing to other benchmarks (Fox et al., 2018; Zaidi et al., 2021), the model size and inference time were over 10 times more efficient; the feedforward computation of the MOTIM was only thousands-level MAC operations, while others were counted in billions. Chapter 6 shifted to postprandial abnormal-glycemia prediction and proposed a novel formulation of this problem that targeted prediction on the fly. The illustrative method proposed for this aim, a joint model using a single LSTM backbone and two linear heads for predicting hyperglycemia and hypoglycemia, outperformed existing benchmarks (Oviedo, Contreras et al., 2019; Seo, Lee et al., 2019; Vehí et al., 2020) by its 0.62 MCC for hyperglycemia and 0.48 MCC for hypoglycemia. In Chapter 7, which concerned the different implementation options observed in existing works, the discussion covered two engineering components, prediction scheme and prediction target, each with two different options. More specifically, the discussed prediction schemes varied in rolling prediction (De Bois, Yacoubi et al., 2019; Zaidi et al., 2021) and direct prediction (Fox et al., 2018; Kim et al., 2020; Rabby et al., 2021; Koca et al., 2023), and the prediction targets vary in single-point prediction (Dong et al., 2019; Li, Liu et al., 2019; Mirshekarian et al., 2019) and excursion prediction (Calm, Garcia-Jaramillo et al., 2007; Adelberger et al., 2021). The experiments evaluating both predictive performance and the efficient suggested that direct prediction might be a better option compared to rolling prediction, while excursion prediction would be beneficial as it provided more insight without loss of efficiency.

In summary, the comparative analysis in this section highlights the contributions of this thesis in advancing data-driven glucose prediction. By integrating novel techniques such as the RSAN and MOTIM, this research not only achieves state-of-the-art results but also offers more efficient and adaptable models compared to the traditional methods in the field.

8.3 Limitations

While this thesis has made contributions in advancing data-driven glucose prediction, it is important to acknowledge the limitations that accompany these findings. Understanding these limitations provides a framework for future improvements and highlights areas that require further exploration.

1. **Data Limitation**: The available datasets, such as the OhioT1DM and UMT1DM datasets, although comprehensive, are limited in size and diversity. These datasets predominantly consist of data from a relatively small number of patients over specific periods, which may not fully capture the variability seen in broader populations. Larger and more varied datasets, encompassing a wider range of demographics, lifestyles, and geographical locations, could provide a more robust validation of the proposed models and ensure their generalisability across different patient groups.

2. **Scenario Specificity**: While the models developed in this thesis were tailored to the specific postprandial scenario, other critical scenarios require further investigation. For instance, the distinction of diurnal and nocturnal prediction, which holds potential to improve the predictive performance, has not been addressed by this thesis.

3. **Exogenous Inputs**: The models primarily rely on CGM data and may not utilise exogenous inputs such as insulin dosages, dietary intake, physical activity, and other lifestyle factors. These exogenous variables can have significant impacts on glucose levels and incorporating them could enhance the predictive accuracy and robustness

of the models. However, accurately and consistently collecting such data can be challenging, and there may be variability in how patients report and log these activities. Future work should focus on integrating these external factors to develop more holistic and accurate glucose prediction models.

Despite the significant advancements made in this thesis, several limitations are identified, including the limited generalisability of the models to diverse patient populations, the constraints posed by the availability and quality of data, and the challenges in integrating real-world factors, highlighting the need for continued efforts to address these issues in future studies.

8.4 Future Works

Building on the findings and limitations identified in this thesis, this section highlights several promising directions for future research. Addressing these areas can further enhance the effectiveness and applicability of data-driven glucose prediction models, contributing to better diabetes management for patients with T1DM.

1. **Multi-modal Approaches**: Integrating additional data modalities, such as insulin dosing, physical activity, and dietary intake, can provide a more comprehensive understanding of glucose dynamics. These multi-modal models can capture the complex interplay between various factors influencing glucose levels, leading to more accurate and insightful predictions.
2. **Broader Scenario Coverage**: Extending the research to cover other critical scenarios, such as nocturnal hypoglycemia prediction and long-term glucose trend analysis, can provide a more holistic approach to diabetes management. By addressing a broader range of scenarios, future work can ensure that glucose prediction models are comprehensive and effective in various contexts.
3. **Real-time Implementation**: Exploring the real-time deployment of these models in wearable devices and smartphone applications, focusing on user-friendly interfaces and real-time feedback mechanisms, can enhance the practical usability of these models. This would require generalising the scope in this thesis to Internet of Things (IoT) engineering.

Future research could focus on addressing the above promising directions, such as improving model generalisability, leveraging larger and more diverse datasets, and incorporating additional real-world factors, to further enhance the robustness and applicability of data-driven glucose prediction models for effective T1DM management.

8.5 Broader Impact

This section analyses the broader implications of research in this thesis for the field of data-driven glucose prediction and beyond, summarised as follows.

- **For Data-driven Glucose Prediction**: This thesis has made the first attempt of using the self-attention mechanism and the meta-optimised time-index modeling for data-driven glucose prediction. This could open the gate of new directions for improving the accuracy and robustness of predictive models in this field, potentially leading to better glucose control and improved quality of life for patients.

- **For Automated Insulin Delivery Systems**: The integration of ML-enhanced glucose prediction models has significant potential to improve the performance of Automated Insulin Delivery (AID) systems, also known as artificial pancreases. These systems, while performing well, can benefit from more accurate glucose predictions, which would lead to better insulin dosing decisions and tighter glucose control. Advanced models like those discussed in this thesis, which leverage self-attention mechanisms and meta-optimised time-index modeling, could enhance AID systems by offering more robust and precise glucose forecasts, ultimately improving patient outcomes in the management of T1DM.
- **For Other Applications in Data Science**: The methodologies and techniques applied in this thesis, particularly the use of self-attention mechanisms and meta-optimised time-index modeling, can be applied to other applications of time-series forecasting within data science. For instance, these techniques can enhance predictive modeling in areas such as financial forecasting, weather prediction, and other biomedical applications like predicting disease progression or treatment outcomes.
- **For Open Science**: The open-source code releases from this thesis offer valuable resources for researchers in both academia and industry. By sharing the codes, other researchers can more easily reproduce and build upon the findings of this thesis, accelerating the advancement of related research areas.

In summary, the broader impact of this research lies in its potential to facilitate novel perspectives for data-driven glucose prediction, promoting similar applications in other domains of data science, and providing open-source resources that can accelerate further innovations.

8.6 Data-driven Glucose Prediction: Rethinking

The previous chapters have presented several advancements made in this thesis toward improving data-driven glucose prediction from various perspectives. However, before concluding, it is valuable to take a step back and reexamine the broader landscape of this field.

Over the past two decades, ML and DL have been increasingly applied to glucose prediction. Yet, unlike in other domains, these techniques have not led to the transformative breakthroughs observed elsewhere. In fields such as computer vision and natural language processing, ML and DL have driven remarkable progress—enabling real-world applications such as autonomous vehicles that navigate through real-time image recognition and virtual assistants like Siri that process human language with impressive accuracy. By contrast, the progress in data-driven glucose prediction appears to have plateaued, with only marginal improvements in predictive performance and limited practical benefits.

This disparity invites critical reflection: why ML and DL have not revolutionised glucose prediction in the same way they have in other domains?

One possible reason is the intrinsic difficulty of data-driven glucose prediction, which causes the gap between ML and DL research to medical research and limits its development. We can understand this point by comparing to other fields, such as computer vision. Taking image recognition as an example — we all can do image recognition, whether we are ML experts, doctor, or anybody who does not know medicine or ML at all. Anyone can sit in front of a screen and do image recognition with a 95% accuracy. This low barrier to entry has fueled rapid advancements in fields like computer vision and natural language processing. However, only experienced practitioners can give reliable prediction on one's glucose trend

(Bunescu et al., 2013). Thus, the research of data-driven glucose prediction does not easily enjoy this advantage.

Fortunately, there are still reasons to remain optimistic about the future of data-driven glucose prediction. The success of increasing data and model size in building intelligent dialogue agents (e.g., the Generative Pre-trained Transformers) has shown that the adage "With enough thrust, even a brick can fly" may indeed hold true in AI. In data-driven glucose prediction, a dominant portion of current research is based on limited publicly available data, which are nowhere near comparable to the amount of data in the successful fields of computer vision and natural language processing, let alone the noise in the measurements, including both glucose and meal information, and the sparsity of important signals, including medication, stress, menses, pregnancy and more. To this end, we can still remain hopeful that the improved research infrastructure in the future can bring new insights and opportunities to data-driven glucose prediction.

8.7 Summary

With the conduction of five individual studies, this thesis has advanced the field of data-driven glucose prediction for T1DM from the perspectives of modeling, adaptation, and engineering. The innovative modeling methods, including the recurrent self-attention network and the meta-optimised time-index models, have enhanced the predictive accuracy and computational efficiency of existing models. The research has also developed effective instance adaptation techniques using transfer learning and meta-learning, enabling reliable predictions with limited personal data. These findings, along with the emphasis on practical engineering solutions, has contributed to more precise and robust glucose prediction technologies. The open-source releases from this work would further support and accelerate future research and development in this field.

Bibliography

Acuna, E., Aparicio, R. and Palomino, V. (2023). 'Analyzing the Performance of Transformers for the Prediction of the Blood Glucose Level Considering Imputation and Smoothing'. *Big Data and Cognitive Computing*, 7(1), p. 41 (cited on pp. 16, 24).

Adelberger, D., Reiterer, F., Schrangl, P., Ringemann, C., Huschto, T. and Re, L. del (2021). 'Prediction of postprandial glucose excursions in type 1 diabetes using control-oriented process models'. *IFAC-PapersOnLine*, 54(15), pp. 466–471 (cited on pp. 7, 24, 25, 64, 75, 76, 78, 87).

Alfian, G., Syafrudin, M., Rhee, J., Anshari, M., Mustakim, M. and Fahrurrozi, I. (2020). 'Blood glucose prediction model for type 1 diabetes based on extreme gradient boosting'. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 803. IOP Publishing, p. 012012 (cited on p. 48).

Allam, F., Nossai, Z., Gomma, H., Ibrahim, I. and Abdelsalam, M. (2011). 'A recurrent neural network approach for predicting glucose concentration in type-1 diabetic patients'. In: *Engineering Applications of Neural Networks*. Springer, pp. 254–259 (cited on pp. 3, 5, 15, 33).

Altuntaş, Y. (2019). 'Postprandial reactive hypoglycemia'. *Şişli Etfal Hastanesi tıp Bülteni*, 53(3), p. 215 (cited on pp. 65, 67).

American Diabetes Association (2010). 'Diagnosis and classification of diabetes mellitus'. *Diabetes care*, 33(Supplement_1), S62–S69 (cited on p. 1).

American Diabetes Association (2023). *Health Equity and Diabetes Technology: A Study of Access to Continuous Glucose Monitors by Payer, Geography and Race Executive Summary*. https://diabetes.org/sites/default/files/2023-09/ADA-CGM-Utilization-White-Paper-Oct-2022.pdf; accessed 31 March 2024 (cited on p. 1).

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B. and De Freitas, N. (2016). 'Learning to learn by gradient descent by gradient descent'. *Advances in neural information processing systems*, 29 (cited on p. 18).

Armandpour, M., Kidd, B., Du, Y. and Huang, J. Z. (2021). 'Deep personalized glucose level forecasting using attention-based recurrent neural networks'. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cited on p. 15).

Athira, V., Geetha, P., Vinayakumar, R. and Soman, K. (2018). 'Deepairnet: Applying recurrent networks for air quality prediction'. *Procedia computer science*, 132, pp. 1394–1403 (cited on p. 11).

Ba, J. L., Kiros, J. R. and Hinton, G. E. (2016). 'Layer normalization'. *arXiv preprint arXiv:1607.06450* (cited on p. 23).

Bahdanau, D., Cho, K. and Bengio, Y. (2014). 'Neural machine translation by jointly learning to align and translate'. *arXiv preprint arXiv:1409.0473* (cited on pp. 10, 12).

Bai, S., Kolter, J. Z. and Koltun, V. (2018). 'An empirical evaluation of generic convolutional and recurrent networks for sequence modeling'. *arXiv preprint arXiv:1803.01271* (cited on pp. 11, 12, 79).

Bechtle, S., Molchanov, A., Chebotar, Y., Grefenstette, E., Righetti, L., Sukhatme, G. and Meier, F. (2021). 'Meta learning via learned loss'. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 4161–4168 (cited on p. 18).

Bengio, S., Vinyals, O., Jaitly, N. and Shazeer, N. (2015). 'Scheduled sampling for sequence prediction with recurrent neural networks'. *Advances in neural information processing systems*, 28 (cited on pp. 76, 77).

Bengio, Y., Courville, A. and Vincent, P. (2013). 'Representation learning: A review and new perspectives'. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), pp. 1798–1828 (cited on p. 2).

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L. et al. (2022). 'Deep learning for time series forecasting: Tutorial and literature survey'. *ACM Computing Surveys*, 55(6), pp. 1–36 (cited on p. 9).

Bergenstal, R. M., Ahmann, A. J., Bailey, T., Beck, R. W., Bissen, J., Buckingham, B., Deeb, L., Dolin, R. H., Garg, S. K., Goland, R. et al. (2013). 'Recommendations for standardizing glucose reporting and analysis to optimize clinical decision making in diabetes: the Ambulatory Glucose Profile'. *Diabetes Technology Society* (cited on p. 67).

Bergstra, J. and Bengio, Y. (2012). 'Random search for hyper-parameter optimization.' *Journal of machine learning research*, 13(2) (cited on p. 111).

Bertachi, A., Biagi, L., Contreras, I., Luo, N. and Vehí, J. (2018). 'Prediction of Blood Glucose Levels And Nocturnal Hypoglycemia Using Physiological Models and Artificial Neural Networks.' In: *KHD@ IJCAI*, pp. 85–90 (cited on pp. 40, 41, 86).

Bertinetto, L., Henriques, J. F., Torr, P. H. and Vedaldi, A. (2018). 'Meta-learning with differentiable closed-form solvers'. *arXiv preprint arXiv:1805.08136* (cited on p. 49).

Bhargav, S., Kaushik, S., Dutt, V. et al. (2021). 'Temporal Convolutional Networks Involving Multi-Patient Approach for Blood Glucose Level Predictions'. In: *2021 International Conference on Computational Performance Evaluation (ComPE)*. IEEE, pp. 288–294 (cited on p. 16).

Bhimireddy, A., Sinha, P., Oluwalade, B., Gichoya, J. W. and Purkayastha, S. (2020). 'Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks'. In: CEUR Workshop Proceedings (cited on pp. 16, 20).

Box, G. E. and Pierce, D. A. (1970). 'Distribution of residual autocorrelations in autoregressive-integrated moving average time series models'. *Journal of the American statistical Association*, 65(332), pp. 1509–1526 (cited on p. 9).

Brew-Sam, N., Chhabra, M., Parkinson, A., Hannan, K., Brown, E., Pedley, L., Brown, K., Wright, K., Pedley, E., Nolan, C. J. et al. (2021). 'Experiences of young people and their caregivers of using technology to manage type 1 diabetes mellitus: Systematic literature review and narrative synthesis'. *JMIR Diabetes*, 6(1), e20973 (cited on p. 68).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). 'Language models are few-shot learners'. *Advances in neural information processing systems*, 33, pp. 1877–1901 (cited on p. 12).

Bunescu, R., Struble, N., Marling, C., Shubrook, J. and Schwartz, F. (2013). 'Blood glucose level prediction using physiological models and support vector regression'. In: *2013 12th*

International Conference on Machine Learning and Applications. Vol. 1. IEEE, pp. 135–140 (cited on p. 90).

Calm, R., Garcia-Jaramillo, M., Vehi, J., Bondia, J., Tarin, C. and Garcia-Gabin, W. (2007). 'Prediction of glucose excursions under uncertain parameters and food intake in intensive insulin therapy for type 1 diabetes mellitus'. In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 1770–1773 (cited on pp. 7, 25, 75, 76, 78, 87).

Calm, R., García-Jaramillo, M., Bondia, J., Sainz, M. and Vehí, J. (2011). 'Comparison of interval and Monte Carlo simulation for the prediction of postprandial glucose under uncertainty in type 1 diabetes mellitus'. *Computer methods and programs in biomedicine*, 104(3), pp. 325–332 (cited on p. 2).

Canete, J. F. de, Gonzalez-Perez, S. and Ramos-Diaz, J. (2012). 'Artificial neural networks for closed loop control of in silico and ad hoc type 1 diabetes'. *Computer methods and programs in biomedicine*, 106(1), pp. 55–66 (cited on p. 25).

Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., Cui, X., Witbrock, M., Hasegawa-Johnson, M. A. and Huang, T. S. (2017). 'Dilated recurrent neural networks'. *Advances in neural information processing systems*, 30 (cited on p. 10).

Chen, T. and Guestrin, C. (2016). 'Xgboost: A scalable tree boosting system'. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794 (cited on pp. 7, 27).

Chen, W., Wilson, J., Tyree, S., Weinberger, K. and Chen, Y. (2015). 'Compressing neural networks with the hashing trick'. In: *International conference on machine learning*. PMLR, pp. 2285–2294 (cited on p. 22).

Chen, Z., Ma, M., Li, T., Wang, H. and Li, C. (2023). 'Long sequence time-series forecasting with deep learning: A survey'. *Information Fusion*, 97, p. 101819 (cited on p. 9).

Chicco, D. and Jurman, G. (2020). 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation'. *BMC Genomics*, 21(1), pp. 1–13 (cited on p. 52).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014). 'Learning phrase representations using RNN encoder-decoder for statistical machine translation'. *arXiv preprint arXiv:1406.1078* (cited on p. 10).

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014). 'Empirical evaluation of gated recurrent neural networks on sequence modeling'. *arXiv preprint arXiv:1412.3555* (cited on pp. 10, 79).

Cui, R., Daskalaki, E., Hossain, M. Z., Lenskiy, A., Nolan, C. J. and Suominen, H. (2021). 'A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning'. In: *2021 15th International Congress in Nursing Informatics (NI)*. IOS Press, pp. 36–38. DOI: 10.3233/SHTI210657 (cited on pp. xvii, 27, 115).

Cui, R., Daskalaki, E., Nolan, C. J. and Suominen, H. (2024). 'Evaluation of Different Engineering Options for Data-driven Glucose Prediction'. In: *TBD* (cited on pp. xvii, 75).

Cui, R., Hettiarachchi, C., Nolan, C. J., Daskalaki, E. and Suominen, H. (2021). 'Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network'. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. DOI: 10.1109/CBMS52027.2021.00064 (cited on pp. xvii, 7, 12, 16, 33, 70, 71, 75–77, 115).

Cui, R., Nolan, C. J., Daskalaki, E. and Suominen, H. (2023). 'Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes'. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. DOI: 10.1109/EMBC40787.2023.10340094 (cited on pp. xvii, 63, 64, 66, 115).

Cui, R., Qian, T., Peng, P., Daskalaki, E., Chen, J., Guo, X., Sun, H. and Jiang, Y.-G. (2022). 'Video moment retrieval from text queries via single frame annotation'. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1033–1043. DOI: 10.1145/3477495.3532078 (cited on p. xvii).

Cui, R., Suominen, H., Nolan, C. J. and Daskalaki, E. (2024). 'Meta-optimised Time-index Modeling for Glucose Excursion Prediction'. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 1911–1918 (cited on pp. xvii, 43, 44, 47, 115).

Cui, Y., Jia, M., Lin, T.-Y., Song, Y. and Belongie, S. (2019). 'Class-balanced loss based on effective number of samples'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277 (cited on p. 19).

D'Antoni, F., Petrosino, L., Marchetti, A., Bacco, L., Pieralice, S., Vollero, L., Pozzilli, P., Piemonte, V. and Merone, M. (2023). 'Layered meta-learning algorithm for predicting adverse events in type 1 diabetes'. *IEEE Access*, 11, pp. 9074–9094 (cited on pp. 20, 21).

Daenen, S., Sola-Gazagnes, A., M'Bemba, J., Dorange-Breillard, C., Defer, F., Elgrably, F., Larger, E. and Slama, G. (2010). 'Peak-time determination of post-meal glucose excursions in insulin-treated diabetic patients'. *Diabetes & Metabolism*, 36(2), pp. 165–169 (cited on pp. 65, 67).

Das, A., Kong, W., Sen, R. and Zhou, Y. (2023). 'A decoder-only foundation model for time-series forecasting'. *arXiv preprint arXiv:2310.10688* (cited on p. 23).

Daskalaki, E., Parkinson, A., Brew-Sam, N., Hossain, M. Z., O'Neal, D., Nolan, C. J. and Suominen, H. (2022). 'The potential of current noninvasive wearable technology for the monitoring of physiological signals in the management of type 1 diabetes: literature survey'. *Journal of medical Internet research*, 24(4), e28901 (cited on p. 43).

De Bois, M., El Yacoubi, M. A. and Ammi, M. (2021). 'Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people'. *Computer Methods and Programs in Biomedicine*, 199, p. 105874 (cited on pp. 20, 86).

De Bois, M., Yacoubi, M. A. E. and Ammi, M. (2019). 'Prediction-coherent LSTM-based recurrent neural network for safer glucose predictions in diabetic people'. In: *International Conference on Neural Information Processing*. Springer, pp. 510–521 (cited on pp. 5, 7, 25, 75–77, 87).

Deng, H., Chen, W., Shen, Q., Ma, A. J., Yuen, P. C. and Feng, G. (2020). 'Invariant subspace learning for time series data based on dynamic time warping distance'. *Pattern Recognition*, 102, p. 107210 (cited on pp. 20, 86).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). 'Imagenet: A large-scale hierarchical image database'. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255 (cited on p. 19).

Deng, Y., Lu, L., Aponte, L., Angelidi, A. M., Novak, V., Karniadakis, G. E. and Mantzoros, C. S. (2021). 'Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients'. *NPJ Digital Medicine*, 4(1), pp. 1–13 (cited on pp. 15, 20, 64, 70, 71, 86).

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). 'Bert: Pre-training of deep bidirectional transformers for language understanding'. *arXiv preprint arXiv:1810.04805* (cited on pp. 12, 19, 34).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014). 'Decaf: A deep convolutional activation feature for generic visual recognition'. In: *International conference on machine learning*. PMLR, pp. 647–655 (cited on p. 19).

Dong, Y., Wen, R., Li, Z., Zhang, K. and Zhang, L. (2019). 'Clu-RNN: A new RNN based approach to diabetic blood glucose prediction'. In: *2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*. IEEE, pp. 50–55 (cited on pp. 7, 25, 75, 76, 78, 87).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). 'An image is worth 16x16 words: Transformers for image recognition at scale'. *arXiv preprint arXiv:2010.11929* (cited on pp. 13, 23, 34).

Du, S., Li, T., Yang, Y. and Horng, S.-J. (2020). 'Multivariate time series forecasting via attention-based encoder–decoder framework'. *Neurocomputing*, 388, pp. 269–279 (cited on p. 11).

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I. and Abbeel, P. (2016). 'RL2: Fast reinforcement learning via slow reinforcement learning'. *arXiv preprint arXiv:1611.02779* (cited on p. 18).

Dubosson, F., Mordvanyuk, N., López Ibáñez, B. and Schumacher, M. (2017). 'Negative results for the prediction of postprandial hypoglycemias from insulin intakes and carbohydrates: analysis and comparison with simulated data'. In: *© Herrero, P., López, B., Martin, C.(eds).(2017). AID 2017: Proceedings of the 2nd International Workshop on Artificial Intelligence for Diabetes held in conjunction with the 16th Conference on Artificial Intelligence in Medicine (AIME): Vienna, Austria: 24th June 2017, p. 25-29*. Artificial Intelligence for Diabetes (AID), Artificial Intelligence in ... (cited on p. 6).

Efendic, H., Kirchsteiger, H., Freckmann, G. and Re, L. del (2014). 'Short-term prediction of blood glucose concentration using interval probabilistic models'. In: *22nd Mediterranean conference on control and automation*. IEEE, pp. 1494–1499 (cited on p. 25).

El Idrissi, T. and Idri, A. (2020). 'Deep learning for blood glucose prediction: Cnn vs lstm'. In: *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part II 20*. Springer, pp. 379–393 (cited on p. 16).

El Idrissi, T., Idri, A., Kadi, I. and Bakkoury, Z. (2020). 'Strategies of Multi-Step-ahead Forecasting for Blood Glucose Level using LSTM Neural Networks: A Comparative Study.' In: *The 13th International Conference on Health Informatics*, pp. 337–344 (cited on p. 7).

Elman, J. L. (1990). 'Finding structure in time'. *Cognitive science*, 14(2), pp. 179–211 (cited on p. 10).

Elsherbini, A. M., Alsamman, A. M., Elsherbiny, N. M., El-Sherbiny, M., Ahmed, R., Ebrahim, H. A. and Bakkach, J. (2022). 'Decoding diabetes biomarkers and related molecular mechanisms by using machine learning, text mining, and gene expression analysis'. *International Journal of Environmental Research and Public Health*, 19(21), p. 13890 (cited on p. 27).

Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J. and Liu, W. (2021). 'You only look at one sequence: Rethinking transformer in vision through object detection'. *Advances in Neural Information Processing Systems*, 34, pp. 26183–26197 (cited on pp. 13, 34).

Felizardo, V., Garcia, N. M., Pombo, N. and Megdiche, I. (2021). 'Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review'. *Artificial Intelligence in Medicine*, 118, p. 102120 (cited on pp. 2, 4, 6, 7, 9, 44, 45, 60).

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y. and Tu, X. M. (2014). 'Log-transformation and its implications for data analysis.' *Shanghai archives of psychiatry*, 26(2), pp. 105–109 (cited on p. 29).

Finn, C., Abbeel, P. and Levine, S. (2017). 'Model-agnostic meta-learning for fast adaptation of deep networks'. In: *International conference on machine learning*. PMLR, pp. 1126–1135 (cited on pp. 6, 18).

Fisher, A., Rudin, C. and Dominici, F. (2019). 'All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously'. *Journal of Machine Learning Research*, 20(177), pp. 1–81 (cited on pp. 27, 29).

Fong, S., Mohammed, S., Fiaidhi, J. and Kwoh, C. K. (2013). 'Using causality modeling and Fuzzy Lattice Reasoning algorithm for predicting blood glucose'. *Expert systems with applications*, 40(18), pp. 7354–7366 (cited on p. 25).

Fox, I., Ang, L., Jaiswal, M., Pop-Busui, R. and Wiens, J. (2018). 'Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories'. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1387–1395 (cited on pp. 5, 7, 15, 25, 46, 48, 53, 75, 76, 78, 79, 86, 87, 107).

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R. and Pontil, M. (2018). 'Bilevel programming for hyperparameter optimization and meta-learning'. In: *International conference on machine learning*. PMLR, pp. 1568–1577 (cited on p. 18).

Freiburghaus, J., Rizzotti, A. and Albertetti, F. (2020). 'A deep learning approach for blood glucose prediction of type 1 diabetes'. In: *Proceedings of the Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence (ECAI 2020), 29-30 August 2020, Santiago de Compostela, Spain*. Vol. 2675. 29-30 August 2020 (cited on p. 20).

Gardner Jr, E. S. (1985). 'Exponential smoothing: The state of the art'. *Journal of forecasting*, 4(1), pp. 1–28 (cited on p. 9).

Godfrey, L. B. and Gashler, M. S. (2017). 'Neural decomposition of time-series data for effective generalization'. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7), pp. 2973–2985 (cited on pp. 14, 45, 48).

Goodfellow, I., Bengio, Y. and Courville, A. (2016). 'Deep Learning'. In: `http://www.deeplearningbook.org`. MIT Press (cited on pp. 2, 3).

Graves, A. and Graves, A. (2012). 'Long short-term memory'. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45 (cited on p. 10).

Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T. and Snyder, M. (2018). 'Glucotypes reveal new patterns of glucose dysregulation'. *PLoS Biology*, 16(7), e2005143 (cited on p. 59).

Hamzaçebi, C., Akay, D. and Kutay, F. (2009). 'Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting'. *Expert systems with applications*, 36(2), pp. 3839–3844 (cited on pp. 76, 77).

Han, J. M., Ang, Y. Q., Malkawi, A. and Samuelson, H. W. (2021). 'Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements'. *Building and Environment*, 192, p. 107601 (cited on p. 11).

He, K., Zhang, X., Ren, S. and Sun, J. (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cited on pp. 11, 23).

Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F. and Liu, Y. (2020). 'Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station'. *Soft Computing*, 24, pp. 16453–16482 (cited on p. 12).

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006). 'A fast learning algorithm for deep belief nets'. *Neural computation*, 18(7), pp. 1527–1554 (cited on p. 19).

Hochreiter, S. and Schmidhuber, J. (1997). 'Long short-term memory'. *Neural Computation*, 9(8), pp. 1735–1780 (cited on pp. 3, 10, 68, 79).

Hospedales, T., Antoniou, A., Micaelli, P. and Storkey, A. (2021). 'Meta-learning in neural networks: A survey'. *IEEE transactions on pattern analysis and machine intelligence*, 44(9), pp. 5149–5169 (cited on pp. 6, 17, 18, 49).

Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M., Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T. et al. (2004). 'Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes'. *Physiological measurement*, 25(4), p. 905 (cited on pp. 2, 3).

Huang, J.-H., He, R.-H., Yi, L.-Z., Xie, H.-L., Cao, D.-s. and Liang, Y.-Z. (2013). 'Exploring the relationship between 5' AMP-activated protein kinase and markers related to type 2 diabetes mellitus'. *Talanta*, 110, pp. 1–7 (cited on p. 28).

Ioffe, S. and Szegedy, C. (2015). 'Batch normalization: Accelerating deep network training by reducing internal covariate shift'. In: *International conference on machine learning*. pmlr, pp. 448–456 (cited on p. 23).

Jang, E., Gu, S. and Poole, B. (2016). 'Categorical reparameterization with gumbel-softmax'. *arXiv preprint arXiv:1611.01144* (cited on p. 22).

Jelinek, H. F., Stranieri, A., Yatsko, A. and Venkatraman, S. (2016). 'Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis'. *Computers in biology and medicine*, 75, pp. 90–97 (cited on p. 28).

Jensen, M. H., Dethlefsen, C., Vestergaard, P. and Hejlesen, O. (2020). 'Prediction of nocturnal hypoglycemia from continuous glucose monitoring data in people with type 1 diabetes: a proof-of-concept study'. *Journal of diabetes science and technology*, 14(2), pp. 250–256 (cited on p. 6).

Jeon, J., Leimbigler, P. J., Baruah, G., Li, M. H., Fossat, Y. and Whitehead, A. J. (2020). 'Predicting glycaemia in type 1 diabetes patients: experiments in feature engineering and data imputation'. *Journal of healthcare informatics research*, 4(1), pp. 71–90 (cited on p. 24).

Jiang, J. and Zhai, C. (2007). 'Instance Weighting for Domain Adaptation in NLP'. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, pp. 264–271. (cited on p. 18).

Jordan, M. I. (1997). 'Serial order: A parallel distributed processing approach'. In: *Advances in psychology*. Vol. 121. Elsevier, pp. 471–495 (cited on p. 10).

Jordan, M. I. and Mitchell, T. M. (2015). 'Machine learning: Trends, perspectives, and prospects'. *Science*, 349(6245), pp. 255–260 (cited on p. 1).

Karim, R. A., Vassányi, I. and Kósa, I. (2020). 'After-meal blood glucose level prediction using an absorption model for neural network training'. *Computers in Biology and Medicine*, 125, p. 103956 (cited on pp. 24, 63, 64).

Khan, H., Sobki, S. and Khan, S. (2007). 'Association between glycaemic control and serum lipids profile in type 2 diabetic patients: HbA 1c predicts dyslipidaemia'. *Clinical and experimental medicine*, 7, pp. 24–29 (cited on p. 28).

Kim, D.-Y., Choi, D.-S., Kim, J., Chun, S. W., Gil, H.-W., Cho, N.-J., Kang, A. R. and Woo, J. (2020). 'Developing an Individual Glucose Prediction Model Using Recurrent Neural Network'. *Sensors*, 20(22), p. 6460 (cited on pp. 5, 7, 16, 24, 25, 33, 75, 76, 78, 87).

Kingma, D. P. and Ba, J. (2014). 'Adam: A method for stochastic optimization'. *arXiv preprint arXiv:1412.6980* (cited on p. 79).

Kirsch, L., Steenkiste, S. van and Schmidhuber, J. (2019). 'Improving generalization in meta reinforcement learning using learned objectives'. *arXiv preprint arXiv:1910.04098* (cited on p. 18).

Koca, Ö. A., Türköz, A. and Kılıç, V. (2023). 'Stacked GRU-Based Glucose Prediction in Type 1 Diabetes'. *European Journal of Science and Technology*, (52), pp. 80–86 (cited on pp. 7, 25, 75, 76, 78, 87).

Koch, G., Zemel, R., Salakhutdinov, R. et al. (2015). 'Siamese neural networks for one-shot image recognition'. In: *ICML deep learning workshop*. Vol. 2. 1. Lille (cited on p. 18).

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). 'Imagenet classification with deep convolutional neural networks'. *Advances in neural information processing systems*, 25 (cited on pp. 11, 23).

Lai, G., Chang, W.-C., Yang, Y. and Liu, H. (2018). 'Modeling long-and short-term temporal patterns with deep neural networks'. In: *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104 (cited on p. 13).

Lai, H., Huang, H., Keshavjee, K., Guergachi, A. and Gao, X. (2019). 'Predictive models for diabetes mellitus using machine learning techniques'. *BMC endocrine disorders*, 19, pp. 1–9 (cited on p. 27).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R. (2019). 'Albert: A lite bert for self-supervised learning of language representations'. *arXiv preprint arXiv:1909.11942* (cited on p. 19).

Langarica, S., Rodriguez-Fernandez, M., Núñez, F. and Doyle III, F. J. (2023). 'A meta-learning approach to personalized blood glucose prediction in type 1 diabetes'. *Control Engineering Practice*, 135, p. 105498 (cited on pp. 20, 21).

Lara-Benítez, P., Carranza-García, M., Luna-Romera, J. M. and Riquelme, J. C. (2020). 'Temporal convolutional networks applied to energy-related time series forecasting'. *applied sciences*, 10(7), p. 2322 (cited on p. 12).

Lara-Benítez, P., Carranza-García, M. and Riquelme, J. C. (2021). 'An experimental review on deep learning architectures for time series forecasting'. *International journal of neural systems*, 31(03), p. 2130001 (cited on p. 9).

LeCun, Y., Bengio, Y. and Hinton, G. (2015). 'Deep learning'. *nature*, 521(7553), pp. 436–444 (cited on p. 2).

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1989). 'Handwritten digit recognition with a back-propagation network'. *Advances in neural information processing systems*, 2 (cited on p. 1).

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). 'Gradient-based learning applied to document recognition'. *Proceedings of the IEEE*, 86(11), pp. 2278–2324 (cited on p. 11).

Li, K., Daniels, J., Liu, C., Herrero, P. and Georgiou, P. (2019). 'Convolutional recurrent neural networks for glucose prediction'. *IEEE Journal of Biomedical and Health Informatics*, 24(2), pp. 603–613 (cited on pp. 16, 25, 33, 75).

Li, K., Liu, C., Zhu, T., Herrero, P. and Georgiou, P. (2019). 'GluNet: A deep learning framework for accurate glucose forecasting'. *IEEE Journal of Biomedical and Health Informatics*, 24(2), pp. 414–423 (cited on pp. 7, 15, 25, 75, 76, 78, 87).

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X. and Yan, X. (2019). 'Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting'. *Advances in neural information processing systems*, 32 (cited on p. 13).

Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017). 'Focal loss for dense object detection'. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988 (cited on p. 19).

Lin, X., Baweja, H., Kantor, G. and Held, D. (2019). 'Adaptive auxiliary task weighting for reinforcement learning'. *Advances in neural information processing systems*, 32 (cited on p. 18).

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X. and Dustdar, S. (2021). 'Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting'. In: *International conference on learning representations* (cited on pp. 13, 34).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). 'Roberta: A robustly optimized bert pretraining approach'. *arXiv preprint arXiv:1907.11692* (cited on pp. 12, 19, 34).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). 'Swin transformer: Hierarchical vision transformer using shifted windows'. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022 (cited on pp. 13, 34).

Loshchilov, I. and Hutter, F. (2016). 'Sgdr: Stochastic gradient descent with warm restarts'. *arXiv preprint arXiv:1608.03983* (cited on p. 79).

Lu, Y., Gribok, A. V., Ward, W. K. and Reifman, J. (2010). 'The importance of different frequency bands in predicting subcutaneous glucose concentration in type 1 diabetic patients'. *IEEE Transactions on Biomedical Engineering*, 57(8), pp. 1839–1846 (cited on p. 25).

Marling, C. and Bunescu, R. (2020). 'The OhioT1DM dataset for blood glucose level prediction: Update 2020'. In: *CEUR Workshop Proceedings*. Vol. 2675. NIH Public Access, p. 71 (cited on pp. 5, 15, 16, 34, 46, 79, 107).

Martinsson, J., Schliep, A., Eliasson, B., Meijner, C., Persson, S. and Mogren, O. (2018). 'Automatic blood glucose prediction with confidence using recurrent neural networks'. In: *3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ IJCAI-ECAI 2018, 13 July 2018*, pp. 64–68 (cited on pp. 40, 41, 86).

Martinsson, J., Schliep, A., Eliasson, B. and Mogren, O. (2020). 'Blood glucose prediction with variance estimation using recurrent neural networks'. *Journal of Healthcare Informatics Research*, 4(1), pp. 1–18 (cited on pp. 3, 5, 16, 25, 33, 48, 64).

Massey Jr, F. J. (1951). 'The Kolmogorov-Smirnov test for goodness of fit'. *Journal of the American Statistical Association*, 46(253), pp. 68–78 (cited on p. 64).

Medtronic (2024). *Guardian - connect continuous glucose monitoring system*. (visited on 1 January 2024) (cited on p. 44).

Menghani, G. (2023). 'Efficient deep learning: A survey on making deep learning models smaller, faster, and better'. *ACM Computing Surveys*, 55(12), pp. 1–37 (cited on p. 6).

Meyes, R., Lu, M., Puiseau, C. W. de and Meisen, T. (2019). 'Ablation studies in artificial neural networks'. *arXiv preprint arXiv:1901.08644* (cited on p. 79).

Mirshekarian, S., Shen, H., Bunescu, R. and Marling, C. (2019). 'LSTMs and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data'. In: *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, pp. 706–712 (cited on pp. 7, 25, 37, 38, 40, 41, 53, 75, 76, 78, 86, 87).

Mougiakakou, S. G., Prountzou, A., Iliopoulou, D., Nikita, K. S., Vazeou, A. and Bartsocas, C. S. (2006). 'Neural network based glucose-insulin metabolism models for children with type 1 diabetes'. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 3545–3548 (cited on p. 15).

Nair, V. and Hinton, G. E. (2010). 'Rectified linear units improve restricted boltzmann machines'. In: *ICML* (cited on pp. 35, 49, 68).

Nemat, H., Khadem, H., Elliott, J. and Benaissa, M. (2020). 'Data fusion of activity and CGM for predicting blood glucose levels'. In: *Knowledge Discovery in Healthcare Data 2020*. Vol. 2675. CEUR Workshop Proceedings, pp. 120–124 (cited on pp. 40, 41, 86).

Nugaliyadde, A., Somaratne, U. and Wong, K. W. (2019). 'Predicting electricity consumption using deep recurrent neural networks'. *arXiv preprint arXiv:1909.08182* (cited on p. 11).

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016). 'Wavenet: A generative model for raw audio'. *arXiv preprint arXiv:1609.03499* (cited on pp. 11, 16).

Ortiz-Martínez, M., González-González, M., Martagón, A. J., Hlavinka, V., Willson, R. C. and Rito-Palomares, M. (2022). 'Recent developments in biomarkers for diagnosis and screening of type 2 diabetes mellitus'. *Current diabetes reports*, 22(3), pp. 95–115 (cited on p. 27).

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D. and Auli, M. (2019). 'fairseq: A fast, extensible toolkit for sequence modeling'. *arXiv preprint arXiv:1904.01038* (cited on p. 34).

Oviedo, S., Contreras, I., Quirós, C., Giménez, M., Conget, I. and Vehí, J. (2019). 'Risk-based postprandial hypoglycemia forecasting using supervised learning'. *International Journal of Medical Informatics*, 126, pp. 1–8 (cited on pp. 6, 24, 52, 64, 65, 67, 70, 71, 87).

Oviedo, S., Vehí, J., Calm, R. and Armengol, J. (2017). 'A review of personalized blood glucose prediction strategies for T1DM patients'. *International Journal for Numerical Methods in Biomedical Engineering*, 33(6), e2833 (cited on pp. 2, 4, 6, 9, 14, 25, 34, 45, 51, 60).

Pan, S. J. and Yang, Q. (2009). 'A survey on transfer learning'. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), pp. 1345–1359 (cited on pp. 6, 17).

Pascanu, R., Mikolov, T. and Bengio, Y. (2013). 'On the difficulty of training recurrent neural networks'. In: *International conference on machine learning*. Pmlr, pp. 1310–1318 (cited on pp. 7, 10, 34, 76, 77).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019). 'Pytorch: An imperative style, high-performance deep learning library'. *Advances in Neural Information Processing Systems*, 32 (cited on p. 111).

Pennington, J., Socher, R. and Manning, C. D. (2014). 'Glove: Global vectors for word representation'. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (cited on p. 19).

Pérez-Gandía, C., Facchinetti, A., Sparacino, G., Cobelli, C., Gómez, E., Rigla, M., Leiva, A. de and Hernando, M. (2010). 'Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring'. *Diabetes technology & therapeutics*, 12(1), pp. 81–88 (cited on p. 24).

Popkin, B. M., Du, S., Zhai, F. and Zhang, B. (2010). 'Cohort Profile: The China Health and Nutrition Survey—monitoring and understanding socio-economic and health change in China, 1989–2011'. *International journal of epidemiology*, 39(6), pp. 1435–1440 (cited on pp. 27, 28, 107).

Prechelt, L. (2002). 'Early stopping-but when?' In: *Neural Networks: Tricks of the trade*. Springer, pp. 55–69 (cited on p. 24).

Pustozerov, E. A., Tkachuk, A. S., Vasukova, E. A., Anopova, A. D., Kokina, M. A., Gorelova, I. V., Pervunina, T. M., Grineva, E. N. and Popova, P. V. (2020). 'Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus'. *IEEE Access*, 8, pp. 219308–219321 (cited on pp. 24, 64).

Qian, T., Cui, R., Chen, J., Peng, P., Guo, X. and Jiang, Y.-G. (2023). 'Locate before answering: Answer guided question localization for video question answering'. *IEEE Transactions on Multimedia*. DOI: 10.1109/TMM.2023.3323878 (cited on p. xvii).

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G. and Cottrell, G. (2017). 'A dual-stage attention-based recurrent neural network for time series prediction'. *arXiv preprint arXiv:1704.02971* (cited on p. 11).

Rabby, M. F., Tu, Y., Hossen, M. I., Lee, I., Maida, A. S. and Hei, X. (2021). 'Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction'. *BMC Medical Informatics and Decision Making*, 21, pp. 1–15 (cited on pp. 5, 7, 15, 25, 75, 76, 78, 87).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). 'Language models are unsupervised multitask learners'. *OpenAI blog*, 1(8), p. 9 (cited on p. 12).

Raghu, A., Raghu, M., Bengio, S. and Vinyals, O. (2019). 'Rapid learning or feature reuse? towards understanding the effectiveness of maml'. *arXiv preprint arXiv:1909.09157* (cited on p. 21).

Rodbard, D. (2016). 'Continuous glucose monitoring: a review of successes, challenges, and opportunities'. *Diabetes technology & therapeutics*, 18(S2), S2–3 (cited on p. 1).

Rubin-Falcone, H., Fox, I. and Wiens, J. (2020). 'Deep Residual Time-Series Forecasting: Application to Blood Glucose Prediction.' *KDH@ ECAI*, 20, pp. 105–109 (cited on pp. 40, 41).

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). 'Learning representations by back-propagating errors'. *Nature*, 323(6088), pp. 533–536 (cited on p. 50).

Rumelhart, D. E. and McClelland, J. L. (1987). 'Learning Internal Representations by Error Propagation'. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, pp. 318–362 (cited on p. 3).

Sacks, D. B. (2011). 'A1C versus glucose testing: a comparison'. *Diabetes care*, 34(2), p. 518 (cited on p. 28).

Seo, W., Lee, Y.-B., Lee, S., Jin, S.-M. and Park, S.-M. (2019). 'A machine-learning approach to predict postprandial hypoglycemia'. *BMC medical informatics and decision making*, 19(1), pp. 1–13 (cited on pp. 6, 24, 64, 65, 67, 70–72, 87).

Seo, W., Park, S.-W., Kim, N., Jin, S.-M. and Park, S.-M. (2021). 'A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study'. *Computer Methods and Programs in Biomedicine*, 211, p. 106424 (cited on p. 20).

Sergazinov, R., Armandpour, M. and Gaynanova, I. (2023). 'Gluformer: Transformer-based Personalized glucose Forecasting with uncertainty quantification'. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5 (cited on p. 16).

Shachee, S., Latha, H. and Hegde Veena, N. (2022). 'Electrical energy consumption prediction using lstm-rnn'. In: *Evolutionary Computing and Mobile Sustainable Networks: Proceedings of ICECMSN 2021*. Springer, pp. 365–384 (cited on p. 11).

Simonyan, K. and Zisserman, A. (2014). 'Very deep convolutional networks for large-scale image recognition'. *arXiv preprint arXiv:1409.1556* (cited on p. 11).

Sitzmann, V., Martel, J., Bergman, A., Lindell, D. and Wetzstein, G. (2020). 'Implicit neural representations with periodic activation functions'. *Advances in Neural Information Processing Systems*, 33, pp. 7462–7473 (cited on pp. 14, 49).

Smith, L. N. (2017). 'Cyclical learning rates for training neural networks'. In: *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 464–472 (cited on p. 24).

Smyl, S. (2020). 'A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting'. *International Journal of Forecasting*, 36(1), pp. 75–85 (cited on p. 10).

Snell, J., Swersky, K. and Zemel, R. (2017). 'Prototypical networks for few-shot learning'. *Advances in neural information processing systems*, 30 (cited on p. 18).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). 'Dropout: a simple way to prevent neural networks from overfitting'. *The journal of machine learning research*, 15(1), pp. 1929–1958 (cited on p. 24).

Sun, Q., Jankovic, M. V., Bally, L. and Mougiakakou, S. G. (2018). 'Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network'. In: *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE, pp. 1–5 (cited on pp. 33, 53).

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. and Hospedales, T. M. (2018). 'Learning to compare: Relation network for few-shot learning'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208 (cited on p. 18).

Sutskever, I., Vinyals, O. and Le, Q. V. (2014). 'Sequence to sequence learning with neural networks'. *Advances in Neural Information Processing Systems*, 27 (cited on pp. 10, 53).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). 'Going deeper with convolutions'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9 (cited on p. 11).

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. and Ng, R. (2020). 'Fourier features let networks learn high

frequency functions in low dimensional domains'. *Advances in Neural Information Processing Systems*, 33, pp. 7537–7547 (cited on p. 48).

Taylor, S. J. and Letham, B. (2018). 'Forecasting at scale'. *The American Statistician*, 72(1), pp. 37–45 (cited on p. 13).

The International Expert Committee (2009). 'International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes'. *Diabetes care*, 32(7), p. 1327 (cited on p. 28).

Thomsen, H. B., Jakobsen, M. M., Hecht-Pedersen, N., Jensen, M. H. and Kronborg, T. (2023). 'Prediction of Hypoglycemia From Continuous Glucose Monitoring in Insulin-Treated Patients With Type 2 Diabetes Using Transfer Learning on Type 1 Diabetes Data: A Deep Transfer Learning Approach'. *Journal of Diabetes Science and Technology*, p. 19322968231215324 (cited on p. 20).

Vanschoren, J. (2018). 'Meta-learning: A survey'. *arXiv preprint arXiv:1810.03548* (cited on p. 6).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). 'Attention is all you need'. *Advances in Neural Information Processing Systems*, 30 (cited on pp. 5, 7, 12, 23, 33–36, 53).

Vehí, J., Contreras, I., Oviedo, S., Biagi, L. and Bertachi, A. (2020). 'Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning'. *Health Informatics Journal*, 26(1), pp. 703–718 (cited on pp. 6, 24, 52, 64, 65, 67, 70, 71, 87).

Vettoretti, M., Cappon, G., Facchinetti, A. and Sparacino, G. (2020). 'Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors'. *Sensors*, 20(14), p. 3870 (cited on p. 1).

Vijayan, V. V. and Anjali, C. (2015). 'Prediction and diagnosis of diabetes mellitus—A machine learning approach'. In: *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. IEEE, pp. 122–127 (cited on p. 27).

Vilalta, R. and Drissi, Y. (2002). 'A perspective view and survey of meta-learning'. *Artificial intelligence review*, 18, pp. 77–95 (cited on p. 17).

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D. et al. (2016). 'Matching networks for one shot learning'. *Advances in neural information processing systems*, 29 (cited on p. 18).

Vu, L., Kefayati, S., Idé, T., Pavuluri, V., Jackson, G., Latts, L., Zhong, Y., Agrawal, P. and Chang, Y.-C. (2019). 'Predicting nocturnal hypoglycemia from continuous glucose monitoring data with extended prediction horizon'. In: *AMIA Annual Symposium Proceedings*. Vol. 2019. American Medical Informatics Association, p. 874 (cited on p. 6).

Weiss, K., Khoshgoftaar, T. M. and Wang, D. (2016). 'A survey of transfer learning'. *Journal of Big data*, 3(1), pp. 1–40 (cited on p. 6).

Wichrowska, O., Maheswaranathan, N., Hoffman, M. W., Colmenarejo, S. G., Denil, M., Freitas, N. and Sohl-Dickstein, J. (2017). 'Learned optimizers that scale and generalize'. In: *International conference on machine learning*. PMLR, pp. 3751–3760 (cited on p. 18).

Woo, G., Liu, C., Sahoo, D., Kumar, A. and Hoi, S. (2023). 'Learning Deep Time-index Models for Time Series Forecasting'. In: *Proceedings of the 40th International Conference on Machine Learning (ICML)* (cited on pp. 5, 7, 14, 43–45, 47, 48, 50).

World Health Organization (2011). *Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation*. Tech. rep. World Health Organization (cited on p. 28).

World Health Organization (2020). *The Top 10 Causes of Death*. (visited on 1 January 2024) (cited on p. 1).

World Health Organization (2023). *Diabetes*. (visited on 1 January 2024) (cited on p. 1).

World Health Organization and International Diabetes Federation (2006). 'Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation' (cited on p. 28).

Wu, H., Xu, J., Wang, J. and Long, M. (2021). 'Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting'. *Advances in Neural Information Processing Systems*, 34, pp. 22419–22430 (cited on pp. 13, 34).

Xingsan, H., Xia, Y., Tao, Y. and Hongru, L. (2021). 'A deep transfer learning model for personalized blood glucose prediction'. In: *2021 China Automation Congress (CAC)*. IEEE, pp. 2045–2049 (cited on p. 20).

Xu, Q., Qu, L., Xu, C. and Cui, R. (2019). 'Privacy-aware text rewriting'. In: *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 247–257. DOI: 10.18653/v1/W19-8633 (cited on p. xvii).

Yang, T., Yu, X., Ma, N., Wu, R. and Li, H. (2022). 'An autonomous channel deep learning framework for blood glucose prediction'. *Applied Soft Computing*, 120, p. 108636 (cited on pp. 64, 70, 71).

Yang, Y. and Hospedales, T. (2016). 'Deep multi-task representation learning: A tensor factorisation approach'. *arXiv preprint arXiv:1605.06391* (cited on p. 18).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V. (2019). 'Xlnet: Generalized autoregressive pretraining for language understanding'. *Advances in neural information processing systems*, 32 (cited on pp. 19, 34).

Ye, L., Rochan, M., Liu, Z. and Wang, Y. (2019). 'Cross-modal self-attention network for referring image segmentation'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10502–10511 (cited on p. 13).

Yu, F. and Koltun, V. (2015). 'Multi-scale context aggregation by dilated convolutions'. *arXiv preprint arXiv:1511.07122* (cited on p. 12).

Yu, X., Yang, T., Lu, J., Shen, Y., Lu, W., Zhu, W., Bao, Y., Li, H. and Zhou, J. (2021). 'Deep transfer learning: a novel glucose prediction framework for new subjects with type 2 diabetes'. *Complex & Intelligent Systems*, pp. 1–13 (cited on pp. 20, 86).

Zaidi, S. M. A., Chandola, V., Ibrahim, M., Romanski, B., Mastrandrea, L. D. and Singh, T. (2021). 'Multi-step ahead predictive model for blood glucose concentrations of type-1 diabetic patients'. *Scientific Reports*, 11(1), p. 24332 (cited on pp. 7, 25, 53, 75–77, 79, 86, 87).

Zarkogianni, K., Mitsis, K., Litsa, E., Arredondo, M.-T., Fico, G., Fioravanti, A. and Nikita, K. S. (2015). 'Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring'. *Medical & Biological Engineering & Computing*, 53, pp. 1333–1343 (cited on p. 48).

Zhang, J., Liu, C. and Ge, L. (2022). 'Short-term load forecasting model of electric vehicle charging load based on mccnn-tcn'. *Energies*, 15(7), p. 2633 (cited on p. 12).

Zhang, J., He, T., Sra, S. and Jadbabaie, A. (2019). 'Why gradient clipping accelerates training: A theoretical justification for adaptivity'. *arXiv preprint arXiv:1905.11881* (cited on p. 24).

Zhao, W., Gao, Y., Ji, T., Wan, X., Ye, F. and Bai, G. (2019). 'Deep temporal convolutional networks for short-term traffic flow forecasting'. *Ieee Access*, 7, pp. 114496–114507 (cited on p. 12).

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. and Zhang, W. (2021). 'Informer: Beyond efficient transformer for long sequence time-series forecasting'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 12, pp. 11106–11115 (cited on pp. 13, 34).

Zhu, T., Chen, T., Kuangt, L., Zeng, J., Li, K. and Georgiou, P. (2023). 'Edge-Based Temporal Fusion Transformer for Multi-Horizon Blood Glucose Prediction'. In: *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, pp. 1–5 (cited on p. 16).

Zhu, T., Kuang, L., Piao, C., Zeng, J., Li, K. and Georgiou, P. (2024). 'Population-specific glucose prediction in diabetes care with transformer-based deep learning on the edge'. *IEEE Transactions on Biomedical Circuits and Systems* (cited on p. 24).

Zhu, T., Li, K., Chen, J., Herrero, P. and Georgiou, P. (2020). 'Dilated recurrent neural networks for glucose forecasting in type 1 diabetes'. *Journal of Healthcare Informatics Research*, 4(3), pp. 308–324 (cited on pp. 3, 34, 40, 86).

Zhu, T., Li, K., Herrero, P., Chen, J. and Georgiou, P. (2018). 'A Deep Learning Algorithm for Personalized Blood Glucose Prediction.' In: *The 3rd International Workshop on Knowledge Discovery in Healthcare Data@IJCAI-ECAI*, pp. 64–78 (cited on pp. 16, 24, 25, 64).

Zhu, T., Li, K., Herrero, P. and Georgiou, P. (2020). 'Deep learning for diabetes: a systematic review'. *IEEE Journal of Biomedical and Health Informatics*, 25(7), pp. 2744–2757 (cited on p. 9).

Zhu, T., Li, K., Herrero, P. and Georgiou, P. (2022). 'Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning'. *IEEE Transactions on Biomedical Engineering*, 70(1), pp. 193–204 (cited on pp. 3, 20).

Zhu, T., Uduku, C., Li, K., Herrero, P., Oliver, N. and Georgiou, P. (2022). 'Enhancing self-management in type 1 diabetes with wearables and deep learning'. *npj Digital Medicine*, 5(1), p. 78 (cited on p. 20).

Zhu, T., Yao, X., Li, K., Herrero, P. and Georgiou, P. (2020). 'Blood glucose prediction for type 1 diabetes using generative adversarial networks'. In: *CEUR Workshop Proceedings*. Vol. 2675, pp. 90–94 (cited on pp. 40, 41, 86).

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2020). 'A comprehensive survey on transfer learning'. *Proceedings of the IEEE*, 109(1), pp. 43–76 (cited on pp. 6, 17, 18).

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. (2018). 'Predicting diabetes mellitus with machine learning techniques'. *Frontiers in genetics*, 9, p. 515 (cited on p. 27).

Appendix A

Datasets

This appendix provides information of the datasets employed in this thesis for further reference. Ethical approval (2020/411) and (2020/515) were obtained from the Human Research Ethics Committee of the Australian National University to use the de-identified publicly available datasets in this thesis.

A.1 OhioT1DM

The dataset created by the US Ohio University, called OhioT1DM (Marling and Bunescu, 2020), contains eight weeks of data from CGM devices and self-reported insulin treatments and meal events for 12 T1DM patients. I followed the train/test split officially provided by the authors of this dataset in the experiments of all my studies presented in Chapter 4 to 7.

The dataset was acquired from
http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html.

A.2 UMT1DM

The dataset created by the US University of Michigan which this thesis has referred to as UMT1DM (Fox et al., 2018), contains daily CGM data of 40 T1DM patients under a three-year protocol: every three months, patients are required to wear the CGM device for several consecutive days. As a result, the data are of a similar amount to OhioT1DM. Following the authors of the dataset, I used the last 3-month data chunk as the test set in my experiments in Chapter 5. As using all 40 patients would result in excessive experiment volume, I randomly selected 12 patients in this dataset without missing evaluation data (i.e., the last three-month chunk) to align with the experiment using OhioT1DM.

The dataset was acquired from
https://github.com/igfox/multi-output-glucose-forecasting/.

A.3 CHNS

The longitudinal data collected in the China Health and Nutrition Survey (CHNS) (Popkin et al., 2010) targets public health among eight provinces in China from 1989 to 2011. I made use of the fasting blood data in 2009 for the study presented in Chapter 3, which contain 30 fasting blood biomarkers together with gender and age information for 9,549 individuals. A list of abbreviations used in this dataset is provided in Table A.1.

Table A.1: Biomarker abbreviations in the CHNS dataset.

Abbreviation	Biomarker
UREA	Urea (mmol/L)
UA	Uric acid (umol/L)
APO_A	Apolipoprotein A-1 (g/L)
LP_A	Lipoprotein (a) mg/L
HS_CRP	High-sensitivity CRP (mg/L)
CRE	Creatinine (umol/L)
HDL_C	High-density lipoprotein cholesterol(mmol/L)
LDL_C	Low-density lipoprotein cholesterol (mmol/L)
APO_B	Apolipoprotein B (g/L)
MG	Serum magnesium (mmol/l)
FET	Ferritin (ng/mL)
INS	Insulin (μIU/mL-2009) (uU/mL-2015)
HGB	Hemoglobin (g/L)
WBC	White Blood Cell ($\times 10^9$ cells/L)
RBC	Red Blood Cell ($\times 10^6$ cells/mcL)
PLT	Platelet ($\times 10^9$ cells/L)
Glu_field	Blood Glucose - Field (mmol/L)
Y48_2	ALT(U/L) Alanine Aminotransferase - Field
Y48_3	TP(g/L) Total Blood Protein - Field
Y48_4	ALB(g/L) Albumin - Field
Y48_5	TC(mmol/L) Total Cholesterol - Field
Y48_6	TG(mmol/L) Triglycerides - Field
HbA1c	Hemoglubin A1c (%)
TP	Total protein (g/L)
ALB	Albumin (g/L)
GLUCOSE	Blood Glucose, (mmol/L)
TG	Triglyceride (mmol/L)
TC	Total cholesterol (mmol/L)
ALT	Alanine Transaminase (U/L)

TRF	Transferrin (mg/dl)
TRF_R	Soluble Transferrin receptor (mg/L)
CRE_MG	CRE[mg/dL] Creatinine
UA_MG	UA[mg/dl] Uric Acid
UREA_MG	UREA[mg/dL] Urea
MG_MG	MG[mg/dL] Magnesium
TC_MG	TC(mg/dL) Total Cholesterol
Y48_5MG	TC(mg/dL) Total Cholesterol - Field
TG_MG	TG(mg/dL) Triglycerides
Y48_6MG	TG(mg/dL) Triglycerides - Field
HDL_C_MG	HDL-C[mg/dL] High-Density Lipoprotein Cholestero
LDL_C_MG	LDL-C[mg/dL] Low-Density Lipoprotein Cholesterol
Glu_field_MG	GLUCOSE(mg/dL) Blood Glucose
GLUCOSE_MG	GLUCOSE(mg/dL) Blood Glucose
APO_A_MG	Apo-A[mg/dL] Apolipoprotein A-1
APO_B_MG	Apo-B[mg/dL] Apolipoprotein B
TRF_MG	TRANSFERRIN (mg/gl)
Y46_1DL	Hb(g/dL) Hemoglobin - Field

Appendix B

Implementation Details

The coding of studies in this thesis was based on Python programming language, where the study in Chapter 3 used the scikit-learn module and the rest chapters used the PyTorch module (Paszke et al., 2019). Normalisation was applied to the CGM data using mean and standard deviation estimated of the training data as a preprocessing step. The rest of this appendix discusses extra implementation details of these studies individually.

B.1 Chapter 3

This study was based on Python scikit-learn module. For the four model employed in the experiment, logistic regression used a solver of liblinear with l2 penalty, and the stopping tolerance followed the default value of 1e-4. The SVM used a linear kernel with squared Hinge loss, and the cost hyperparameter C was set to 1.0. The MLP contained a single hidden layer of size 100 activated by ReLU, and the Adam optimiser was used to tuned the MLP with default settings of $\alpha = 0.0001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Finally, the xgbooster algorithm employed gbtree as the booster, where the total number of estimators was set to 100.

B.2 Chapter 4

In the study in Chapter 4, the hyperparameters of the training pipeline were determined through a grid search (Bergstra and Bengio, 2012), whereby a 3-layer encoder with the dimension of the hidden representation being 128 was used. As a result, the final RSAN model contained around 0.60 million trainable parameters. The model was trained using Adam optimiser on an NVIDIA V100 GPU with a batch size of 16. Under these settings, each training step took around 50 milliseconds. The learning rate scheduling scheme was halving the learning rate on plateaus with its patience being 1 epoch. With this decaying scheme, I pre-trained the model for 10 epochs with an initial learning rate of 0.0003 and fine-tuned the model for 10 epochs with an initial learning rate of 0.00005. Lastly, the training and testing examples with missing entries were excluded in all experiments in this study.

B.3 Chapter 5

In the data preprocessing, I conducted linear interpolation for missing glucose recordings no longer than 15 minutes (3 samples) in purpose of making the maximal use of the data. In training the models, I applied mean squared error (MSE) between the predicted excursion and the ground truth as the loss function. For the training/validation split, I constantly left out the last 20% of the training data as the validation set. All models were tuned using Adam optimiser with a fixed learning rate of 1e-3 and a batch size of 64. All experiments

Table B.1: Experimental setup details of the Chapter 5 study.

Setup of	Details
MOTIM	layer_size: 128 inr_layers: 1 n_fourier_feats: 256 σ^2 scales: [0.01, 0.1, 1, 5, 10, 20, 50, 100]
MoGRU	layer_size: 512 num_layer: 2 bidirectional: False
ReSelfAttn	layer_size: 128 num_layer: 3 n_heads: 4 d_ff: 512
SeqConv	layer_size: 64 num_layer: 3
OhioT1DM	anchor_patients: [540, 544, 552, 567, 584, 596, 559, 563, 570, 575, 588, 591]
UMT1DM	anchor_patients: [2, 5, 7, 9, 11, 15, 17, 19, 21, 27, 30, 33]
Training	epochs: 50 batch_size: 64 optimiser: Adam learning_rate: 1e-3 gradient_clipping: 10.0 early_stopping_patience: 7 epochs

were conducted on an NVIDIA RTX3090 GPU. For the historical-value benchmarks, I set the input length to 60 minutes and empirically adopted the model specifications in the original studies such as the hidden size and number of layers in our replication. For the MOTIM, I applied a one-layer INR of hidden size 128, which took 256 Fourier features as its input, and conducted a hyperparameter search for the lookback length in [1, 3, 5, 7] times the length of the prediction horizon. A summary of experimental setup details can be found in Table B.1.

B.4 Chapter 6

In the model configuration, I applied a uni-layer LSTM with hidden dimension of 50, resulting in the full model contained around 10k trainable parameters. The model was tuned using Adam optimiser with a batch size of 64 and a learning rate of 0.001 under a gradient norm clipping at 1.0. To illustrate the computational efficiency under this configuration, each training step took around 0.01 seconds, and the full training pipeline for all 11 patients in OhioT1DM dataset took around 4 minutes on an Apple M1 CPU. As the amount of personalised postprandial CGM data was limited, instead of using held-out training data as a validation set and conduct model selection, I trained the model on the full training data and fixed the training epochs to 50.

B.5 Chapter 7

Experiments in this chapter were based on the coding framework of the Chapter 5 study, in which implementations of the LSTM, the GRU, and the TCN were created and merged in. To avoid repetition, readers are referred to the implementation details of Chapter 5.

Appendix C

Publications

The manuscripts of the studies in Chapter 3-6 are available as follows.

Chapter 3

R. Cui, E. Daskalaki, M. Z. Hossain, A. Lenskiy, C. J. Nolan and H. Suominen (2021). 'A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning'. In: *2021 15th International Congress in Nursing Informatics (NI)*. IOS Press, pp. 36–38. DOI: 10.3233/SHTI210657

Chapter 4

R. Cui, C. Hettiarachchi, C. J. Nolan, E. Daskalaki and H. Suominen (2021). 'Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network'. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. DOI: 10.1109/CBMS52027.2021.00064

Chapter 5

R. Cui, H. Suominen, C. J. Nolan and E. Daskalaki (2024). 'Meta-optimised Time-index Modeling for Glucose Excursion Prediction'. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 1911–1918

Chapter 6

R. Cui, C. J. Nolan, E. Daskalaki and H. Suominen (2023). 'Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes'. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. DOI: 10.1109/EMBC40787.2023.10340094

36 Nurses and Midwives in the Digital Age
M. Honey et al. (Eds.)

doi:10.3233/SHTI210657

A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning

Ran CUI[a,1], Elena DASKALAKI[a], Md Zakir HOSSAIN[a], Artem LENSKIY[a], Christopher J NOLAN[b] and Hanna SUOMINEN[a,c,d]
[a] *School of Computing, The Australian National University (ANU), Australia*
[b] *ANU Medical School and John Curtin School of Medical Research, ANU, Australia*
[c] *Data61, Commonwealth Scientific and Industrial Research Organisation, Australia*
[d] *Department of Computing, University of Turku, Finland*

Abstract. Diabetes can be diagnosed by either Fasting Plasma Glucose or Hemoglobin A1c. The aim of our study was to explore the differences between the two criteria through the development of a machine learning based diabetes diagnostic algorithm and analysing the predictive contribution of each input biomarker. Our study concludes that fasting insulin is predictive of diabetes defined by FPG, but not by HbA1c. Besides, 28 other fasting blood biomarkers were not significant predictors of diabetes.

Keywords. Diabetes biomarkers, machine learning, feature importance

1. Introduction

Diabetes is a chronic disease which affects a large and increasing number of people worldwide. According to the *World Health Organisation* (WHO), diabetes can be diagnosed by either a *Fasting Plasma Glucose* (FPG) ≥ 7.0 mmol/l or a HbA1c ≥ 6.5%. Although FPG and HbA1c exhibit a proven correlation [1], the two diagnostic criteria FPG ≥ 7.0 mmol/l and HbA1c ≥ 6.5% are not always equivalent leading often to contradicting outcomes [2]. The aim of our study was to explore the difference of the two aforementioned diabetes diagnostic criteria through *Machine Learning* (ML) as well as investigating the contribution of 30 separate blood biomarkers[2] on diabetes prediction.

2. Approach

After obtaining ethical approval (2020/515) from the ANU Human Research Ethics Committee, we analysed longitudinal data collected in the *China Health and Nutrition Survey* (CHNS) [3], which targeted public health among eight provinces in China from 1989 to 2011. We made use of the 30 fasting blood biomarkers included in the CHNS,

[1] Corresponding Author, Ran Cui; E-mail: ran.cui@anu.edu.au.

together with gender and age information of 9,549 individuals. Among all these participants, 550 and 642 subjects satisfied FPG ≥ 7.0 mmol/l and HbA1c ≥ 6.5%, respectively. We formulated the diagnosis of diabetes as a ML classification task tackled with the XGBoost algorithm, which took the biomarkers, gender, and age as inputs. We applied stratified sampling to balance the number of positive (diabetes) and negative (non-diabetes) samples in the experiments. To better investigate the relationship between FPG and HbA1c, we designed the following two experiments: 1) excluding both FPG and HbA1c in the input features set, while using FPG ≥ 7.0 mmol/l or HbA1c ≥ 6.5% as the definition of diabetes in separate analyses (experiment 1), and including FPG or HbA1c in the input features set when the other was used as the ground truth label (experiment 2). We then employed permutation feature importance [4] to study the contribution of each feature (biomarker) towards the final prediction of diabetes diagnosis.

3. Results

Performance, measured by the F1 score, dropped when both FPG and HbA1c were removed from the input biomarkers set (Table 1). This agrees with the aforementioned proven correlation of FPG and HbA1c and their utility for diabetes diagnosis. Although fasting insulin (INS) was also an important predictive biomarker (Table 2) when diabetes was diagnosed by FPG ≥ 7.0 mmol/l, it was not predictive of diabetes diagnosed by HbA1c. Moreover, age possessed high importance in our models, which showed its strong position as a diabetes risk factor. None of the remaining biomarkers were strongly correlated with the presence of diabetes, which suggested that they were not informative and can be safely discarded from our model without affecting algorithmic performance.

Table 1. F1 scores in our experiments.

Diabetes definition	Experiment 1	Experiment 2
FPG ≥ 7.0 mmol/l	0.822	0.893
HbA1c ≥ 6.5%	0.791	0.867

Table 2. Sorted average permutation features importance [%] in the two experiments. Features (biomarkers) that were significant have been bolded.

Diabetes Definition	Experiment 1		Experiment 2		Diabetes Definition	Experiment 1		Experiment 2	
FPG ≥ 7.0 mmol/	**INS**	**2.85**	**HbA1c**	**6.96**	HbA1C ≥ 6.5%	**age**	**1.40**	**FPG**	**7.04**
	age	**1.42**	**INS**	**3.32**		MG	0.59	ALT	0.31
	ALB	0.57	age	0.40		APO_B	0.48	TP	0.29
	UA	0.52	Y48_3	0.28		HS_CRP	0.41	APO_A	0.29
	Y48_4	0.45	APO_A	0.17		APO_A	0.37	Age	0.28
	MG	0.39	Y48_4	0.13		FET	0.34	RBC	0.28
	…	…	…	…		…	…	…	…
	TRF_R	-0.05	Y48_5	-0.22		LP_A	-0.27	UA	-0.03
	LDL_C	-0.07	Y48_2	-0.22		TRF	-0.35	PLT	-0.09

4. Conclusions

We conclude our study with two clinical suggestions: (1) Our finding of fasting insulin being an important biomarker only when diabetes is defined by FPG, but not by HbA1c, suggests that HbA1c should be considered in conjunction with insulin levels in fasting blood samples to ensure an accurate diagnosis, while this does not hold if the diagnosis is based on FPG. (2) Most biomarkers other than FPG, HbA1c, and fasting insulin included in our experiments were not significantly more important than age, hence a clinician could refer less to them when predicting diagnosis of diabetes.

Acknowledgements

This research has been delivered in partnership with and funded by *Our Health in Our Hands* (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. We acknowledge the funding from the ANU School of Computing for the first author's PhD studies.

References

[1] Khan H, Sobki S, Khan S. Association between glycaemic control and serum lipids profile in type 2 diabetic patients: HbA1c predicts dyslipidaemia. Clinical and Experimental Medicine 7 (2007), 24–29.
[2] Sacks DB. A1C versus glucose testing: a comparison. Diabetes Care 34 (2011), 518–523.
[3] Popkin BM, Du S, Zhai F, Zhang B. Cohort Profile: The China Health and Nutrition Survey— monitoring and understanding socio-economic and health change in China, 1989–2011. International Journal of Epidemiology 39 (2009), 1435–1440.
[4] Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. arXiv, arXiv preprint arXiv:1801.01489 (2018).

Personalised Short-Term Glucose Prediction via Recurrent Self-Attention Network

Ran Cui*, Chirath Hettiarachchi*, Christopher J Nolan†, Elena Daskalaki* and Hanna Suominen‡
*‡ School of Computing, The Australian National University, Canberra, Australia
†Medical School and The John Curtin School of Medical Research, The Australian National University, Canberra, Australia
‡Data61, Commonwealth Scientific and Industrial Research Organisation, Canberra, Australia
‡Department of Computing, University of Turku, Turku, Finland
Emails: {ran.cui, chirath.hettiarachchi, christopher.nolan, eleni.daskalaki, hanna.suominen}@anu.edu.au

***Abstract*—People with type 1 diabetes mellitus (T1DM) must continuously monitor their blood glucose levels and regulate them by insulin dosing to stay in a safe range. A reliable glucose prediction technique could pre-alert the risk of abnormal glycaemia in the near future. In this paper, we apply an attention-based deep network for glucose prediction, which models both the temporal dependencies and the physiological relations among glucose and glucose-related life events, including insulin and carbohydrate intake. We also propose to make use of the knowledge learned from non-target subjects with the same disease to improve the personalised prediction by applying parameters transfer. Our approach was evaluated on the standard benchmark OhioT1DM dataset, where the experiments achieved average root mean square errors over the 12 subjects of 17.82 mg/dL for 30 minutes and 28.54 mg/dL for 60 minutes. Additionally, our ablation experiments indicated that the use of transfer learning constantly improved the prediction. On this basis, we conclude that our approach achieves state-of-the-art performance with statistical significance, and data from other people with T1DM could help on improving personalised predictions. We release our codes at https://github.com/r-cui/GluPred under the MIT license.**

***Index Terms*—glucose prediction, time-series, deep learning**

I. Introduction

T1DM is an autoimmune condition in which the immune system attacks pancreatic beta cells, resulting in a deficiency of insulin secretion. Therefore, people with T1DM often require exogenous insulin to maintain their blood glucose in a safe range. Accurate prediction of blood glucose would enable people with T1DM to be proactive on any potential dangerous hyperglycaemia or hypoglycaemia events soon. Traditional glucose prediction methods tend to use physiological models that model the glucose metabolism directly, such as the Dalla Man Model [1] and the Hovorka model [2]. With the emergence of Continuous Glucose Monitoring (CGM) systems, large amounts of continuous glucose records together with data of other glucose-related factors such as insulin intake are becoming available [3]. This enables the great potential of glucose prediction using deep learning based data-driven methods, which enjoy the advantage of extensibility and personalisability.

The key to a successful deep learning based glucose prediction is learning both the temporal dependency and the dynamic relation among features. Because of the temporal nature of the data, the most common trend is using Recurrent Neural Networks (RNNs) [4]–[7]. However, these methods suffer from several pitfalls. RNNs are long known for their training difficulties due to vanishing and exploding gradients [8]. Furthermore, the long-term history forgetting problem [9] remains a drawback of fully recurrent deep networks, though variant versions of RNN such as Long Short-Term Memory (LSTM) [10] have mitigated it. As a result, the prediction from RNNs tends to be less sensitive to early insulin bolus infusion events and food intake events.

We therefore apply a different approach to encode the temporal information of CGM data. Inspired by the success of Transformer [11] models on various natural language and computer vision tasks [12]–[15], we make use of their self-attention mechanism to encode the temporal structure of CGM data. Specifically, the model learns to distribute its "attention" across all timestamps in the input, that is, the model assigns each timestamp a weight that sums to one given the current predictive circumstance every time at the beginning of making a prediction. The model then encodes the input as a linear combination of the deep representation of all timestamps weighted by the attention values. Because the self-attention mechanism is feed-forward, we minimise the recurrence in the network and keep the inevitable step-by-step prediction process as the only recurrence in our deep learning pipeline, thus alleviating the aforementioned problems of RNNs.

Additionally, inspired by the fact that people diagnosed with T1DM have similar pathogenesis, and also inspired by the positive results of previous studies [16]–[18] that predict glucose using pure physiological models which only have a minimum number of parameters to fit, we argue that using data from other people with T1DM could support our objective of personalised glucose prediction. To this end, we adopt the parameter transfer technique [19] to let our model first learn the generic glucose-insulin-carbohydrate metabolism from data of other subjects with T1DM. After this pre-training stage, we fine-tune the model parameters using data of the target subject to continue learning the personalised patterns.

In summary, our contributions are: (i) we apply a new feed-forward model that encodes both the temporal information and the inter-feature relations to short-term glucose prediction; (ii) we give evidence that data from other subjects with T1DM could help the learning of a personalised predictive system.

II. Related Work

As reviewed in [20], modern methods for glucose prediction can be classified into the following two categories: (i) physiological models, which explicitly model the metabolism of glucose and insulin, and (ii) data-driven models, which predict glucose merely by analysing the patterns in CGM data. Since our study belongs to the second category, we only focus on the recent developments in data-driven models.

Data-driven short-term glucose prediction was treated as a simple uni-modal regression problem in the early stage. As first discussed in [21], experiments suggested that large amounts of data are the necessity to enable reliable prediction. After that, a series of methods were applied on unimodal CGM data, such as the first-order autoregressive model and its several variations [22], [23], meta-learning [24], and fully connected feed-forward neural networks [25]. These studies indicated that both statistical models and machine learning based methods could be applied to this problem.

On the other hand, the modality of data used in glucose prediction research was generalised from unimodal CGM data to CGM data with other covariates like insulin therapy and food intake. Namely, experiments indicated that predicting glucose with only past glucose values comes with limitations [26], [27], with the most significant one being that the predictor is insensitive to the sudden changes of glucose values, including both intrinsic changes (i.e., individual metabolic fluctuations) and extrinsic changes (i.e., meal intake or insulin injections). Therefore, the research was pointed to a direction that takes into account multiple factors in addition to glucose.

A significant problem of the research on glucose prediction was that different studies were often weakly connected, that is, researchers tended to use their own collected data and methods for evaluation. Due to this inconsistency on datasets and evaluation metrics, it was difficult to intuitively compare the performance of different methods. The release of the OhioT1DM dataset [28] in 2018 provided pre-defined training/testing data, thus establishing a good benchmark for evaluating across different methods. On the other hand, deep learning techniques have been shown to be effective in personalised glucose prediction. Due to the sequential structure of CGM data, RNNs and one-dimensional Convolutional Neural Networks (CNNs) have become the common models for this task [7], [29]. On this basis, variations of RNN and CNN were evaluated and some, such as causal convolution with dilated kernels [30] and dilated RNN [31] were shown to be better performing. The OhioT1DM dataset was augmented in 2020 [32] with a new round of data collection of approximately the same amounts of data as in the first release. At the moment, LSTM [33] and Generative Adversarial Network (GAN) [34] are leading the state of the art on the two sub-datasets, respectively.

Our above analysis of existing literature indicates that research in data-driven glucose prediction should make use of multi-modal data, and should be evaluated on open datasets to enable a quantitative comparison, both of which are targeted by our paper.

III. Proposed Approach

We define the glucose prediction problem based on a general sequence prediction problem. Given the input sequence $\mathbf{X}_{1:N} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ where $\mathbf{x}_i \in \mathbb{R}^F$ with F being the number of features of the sequence. The target is to predict the future sequence $\mathbf{X}_{N+1:N+T}$ where T is referred to as the prediction horizon. When mapping this definition to glucose prediction, the sequence features consist of glucose and other covariants, such as the insulin and carbohydrate intake information. To this end, we propose a *recurrent self-attention network* that encodes the temporal dependencies, models the complex inter-feature dynamic relations, and makes predictions accordingly.

A. Self-Attention Layer

In glucose prediction, insulin bolus and carbohydrates intake events are naturally sparse, yet essential for making accurate predictions due to their direct effects on glucose concentration. To properly capture this property in our temporal data, we make use of the self-attention mechanism [11]. With this mechanism, the model is able to focus on the positions that are most relevant to the learning objective (Fig. 1).

To formally define the concepts, an i-th *self-attention layer* takes as input a matrix $\mathbf{H}^{(i)} \in \mathbb{R}^{N\times D}$ where N is the length of the input sequence and D is the hidden dimension. The input $\mathbf{H}^{(i)}$ is then linearly transformed to get the query $\mathbf{Q}^{(i)} \in \mathbb{R}^{N\times D}$, key $\mathbf{K}^{(i)} \in \mathbb{R}^{N\times D}$, and value $\mathbf{V}^{(i)} \in \mathbb{R}^{N\times D}$ by

$$\mathbf{Q}^{(i)} = \mathbf{H}^{(i)}\mathbf{W}_q^{(i)}, \tag{1}$$

$$\mathbf{K}^{(i)} = \mathbf{H}^{(i)}\mathbf{W}_k^{(i)}, \text{ and} \tag{2}$$

$$\mathbf{V}^{(i)} = \mathbf{H}^{(i)}\mathbf{W}_v^{(i)} \tag{3}$$

where each of $\mathbf{W}_q^{(i)}, \mathbf{W}_k^{(i)}$, and $\mathbf{W}_v^{(i)}$ is in $\mathbb{R}^{D\times D}$. The self-attention $\mathbf{A}^{(i)} \in \mathbb{R}^{N\times D}$ is then defined as

$$\mathbf{A}^{(i)} = \text{softmax}(\frac{\mathbf{Q}^{(i)}{\mathbf{K}^{(i)}}^\top}{\sqrt{D}})\mathbf{V}^{(i)}. \tag{4}$$

The self-attention layer outputs a nonlinear transformation $\sigma^{(i)}(\cdot)$ of the self-attention $\mathbf{A}^{(i)}$. Here, we apply a fully-connected feed-forward network on the hidden dimension D with a ReLU activation as the nonlinearity $\sigma^{(i)}(\cdot)$. The output of a self-attention layer is consequently defined as

$$\mathbf{H}^{(i+1)} = \sigma^{(i)}(\mathbf{A}^{(i)}) \tag{5}$$

$$= \text{ReLU}(\mathbf{A}^{(i)}\mathbf{W}_1^{(i)} + \mathbf{b}_1^{(i)})\mathbf{W}_2^{(i)} + \mathbf{b}_2^{(i)} \tag{6}$$

where the output $\mathbf{H}^{(i+1)}$ remains in $\mathbb{R}^{N\times D}$.

B. Recurrent Self-Attention Network

To increase the capacity of the model, we stack multiple self-attention layers to form the *encoder* of our *self-attention network*. Besides the encoder, the network makes use of two trainable linear transformations to bridge data between spaces with different dimensions. Specifically, one linear operation is applied to map the input from low-dimensional feature space

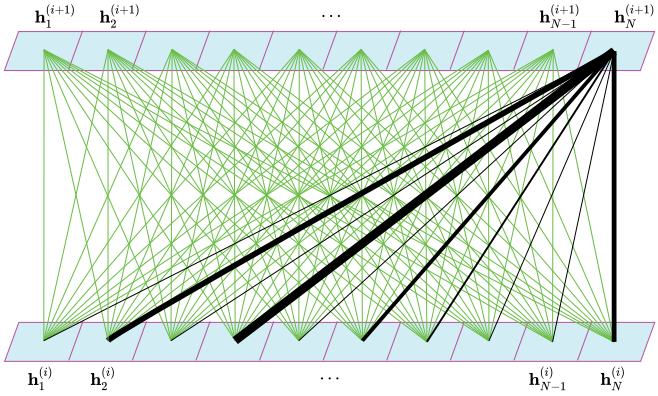


Fig. 1. A high-level illustration of a self-attention layer. The core concept in (4) is a dense connection among all elements in the input and output sequence. Specifically, each element $\mathbf{h}_j^{(i+1)}$ in the output sequence has N weights on all the elements in the input sequence that sums to one. To this end, the model is capable to selectively distribute its attention on the most relevant positions of the input sequence, as illustrated in the example of $\mathbf{h}_N^{(i+1)}$ where the widths of black lines reveal the attention weights.

$\mathbb{R}^F$ to high-dimensional hidden space $\mathbb{R}^D$ before data enter the encoder; the other linear operation contrarily maps the output of the encoder back to the prediction space $\mathbb{R}^F$. Apart from this, given the fact that the aforementioned transformations are all w.r.t. the hidden dimension and identical to each timestamp in the sequence, our self-attention mechanism is invariant to the order of the sequence. To enable the model learning the sequential patterns, we apply sinusoidal positional encoding [11] before the first layer of the encoder. To summarise, the self-attention network takes as input a matrix $\mathbf{X} \in \mathbb{R}^{N\times F}$ and outputs a matrix $\mathbf{Y} \in \mathbb{R}^{N\times F}$, where $\mathbf{y}_N \in \mathbb{R}^F$ is the prediction of $\mathbf{x}_{N+1}$.

To predict T steps ahead, we augment our defined concepts to form the recurrent self-attention network. The recurrent self-attention network is a serial connection of T self-attention networks, where each latter self-attention network takes as input the concatenation of the input and the one-step prediction of the former self-attention network. Following the standard practice of recurrent deep networks, we make the T networks sharing their weights. Since all the operations are differentiable w.r.t. the parameters, the full network can be trained using standard back-propagation.

C. Transfer Learning

Apart from the model structure, we employ transfer learning as part of our training pipeline to further increase the performance of our predictive system. Consequently, the problem statement is generalised from a basic glucose prediction problem defined at the beginning of Section III to a joint learning scenario of M personalised predictive systems for M different subjects with T1DM. According to the taxonomy in [19], this scenario is referred to as multi-task learning because both source and target domain labels are available. To this end, we employ the parameter-transfer technique, where we split the training process into a pre-training stage on data from non-target subjects, and a fine-tuning stage on personal data of the target subject. With this technique, the model reserves the knowledge of generic glucose metabolism learned from non-target subjects, then makes use of the knowledge to better converge itself on an appropriate optimum for the target subject.

IV. Experiments

A. Datasets

We experimented with our approach on both the 2018 and the 2020 releases of the OhioT1DM dataset [32] which contain eight weeks of data from CGM devices, physiological sensors and self-reported life events for 12 people with T1DM. It provides a pre-defined train/test split for each subject, where the testing set comprises around ten days of data, thus enabling a fair comparison of our approach against existing methods. In the dataset, glucose is recorded every 5 minutes. Therefore, we manually aligned the other features to the glucose recordings. Besides, the glucose records in the dataset are not fully continuous with missing entries of different time periods. To maximise the validity of our experiments, we excluded any training and testing examples with missing entries. Ethical approval (2020/411) was obtained from the Human Research Ethics Committee of The Australian National University to use de-identified data from the OhioT1DM dataset for this study.

B. Experiment Setups

We included the interstitial fluid glucose concentration from the CGM recordings, basal insulin infusion rate, insulin bolus delivery and carbohydrate oral intake in our features set as they are commonly considered as having the most direct physiological effects on glucose changes, resulting in $F = 4$. Moreover, two hours of historical data were used to predict the glucose level at 30 minutes and 60 minutes in the future, respectively, following the common setups in the literature.

To bring the research close to real-life applications, we experimented with our approach in two scenarios, inspired by [33].

Glucose Prediction (Setting 1): In this setting, the model knew the ground truth of non-glucose data till the moment of each prediction step. Namely, the basal rate, and any bolus or carbohydrate intake events in the prediction horizon were directly provided to the model in both the training and testing stages. This "what-if" scenario answered questions of what the glucose level will be if any glucose-related event in the features set is planned to happen within the prediction horizon.

All Features Prediction (Setting 2): In contrast to setting 1, the model was not provided with any information in the prediction horizon for this scenario. This implied that the model needed to implicitly learn the life patterns of the subject, and predict the life events together with the glucose values to have a performance comparable to setting 1. A trained model under this scenario could serve as a fully diagnostic tool of what the glucose level will most likely be according to the two-hour historical information only.

TABLE I
RESULTS OF 30 MINUTES PREDICTION HORIZON.[1,2,3]

		RMSE (mg/dL)												
Subject		*540*	*544*	*552*	*567*	*584*	*596*	Avg (SD)	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
LSTM [7]					-				19.5	19.0	16.5	24.2	19.2	22.0	20.07 (2.67)
Feed-Forward [35]					-				18.83	19.43	15.88	22.86	17.84	21.12	19.33 (2.45)
Dilated RNN [31]					-				18.6	**18.0**	15.3	22.7	17.6	21.1	18.88 (2.64)
LSTM [33]	Set. 1				-				individual RMSEs unprovided						18.19
	Set. 2														18.74
Regression [36]		20.98	17.66	16.30	20.52	21.62	17.45	19.09 (2.22)	-						
GAN [34]		20.14	16.28	16.08	20.00	20.91	16.63	18.34 (2.23)	-						
Our approach	Set. 1	**19.49**	15.75	15.68	**18.96**	**19.52**	15.94	**17.56** (1.59)*	**17.57**	18.34	14.84	21.69	16.02	20.08	18.09 (2.53)*
	Set. 2	20.24	**15.26**	**15.49**	19.72	20.23	**15.90**	17.81 (2.49)*	17.59	18.34	**14.64**	**21.31**	**15.93**	**20.03**	**17.97** (2.49)*

TABLE II
RESULTS OF 60 MINUTES PREDICTION HORIZON.[1,2,3]

		RMSE (mg/dL)												
Subject		*540*	*544*	*552*	*567*	*584*	*596*	Avg (SD)	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
LSTM [7]					-				34.4	29.9	28.6	37.3	33.1	36.0	33.22 (3.41)
Feed-Forward [35]					-				32.52	31.33	27.48	35.28	30.12	33.60	31.72 (2.74)
LSTM [33]	Set. 1				-				individual RMSEs unprovided						29.12
	Set. 2														30.63
Regression [36]		39.05	30.42	29.38	36.52	37.01	28.92	33.55 (4.46)	-						
GAN [34]		38.54	27.64	29.03	35.65	34.31	28.10	32.21 (4.57)	-						
Our approach	Set. 1	**32.79**	**25.23**	**25.71**	**30.87**	**31.20**	**25.14**	**28.49** (3.50)*	29.56	**28.37**	**25.04**	**32.34**	**26.73**	**29.56**	**28.60** (2.53)*
	Set. 2	38.74	26.57	28.56	34.49	31.48	26.23	31.01 (4.91)*	**29.31**	29.82	26.22	32.61	27.61	31.66	29.54 (2.40)*

C. Implementation Details

The full deep learning pipeline was implemented using PyTorch. We followed a multi-head attention implementation with 4 heads instead of a single attention function, as suggested by [11]. A grid search was performed to find a good set of hyperparameters. For the model structure, we used a 3-layer encoder, with hidden dimensions being 128 for the multi-head attention modules and 512 for the feed-forward modules. As a result, our model contained around 0.60M trainable parameters. We trained the model using Adam optimiser on an NVIDIA V100 GPU with batch size 16. Under these settings, each training step took around 50 milliseconds. Our learning rate scheduling scheme was halving the learning rate on plateaus with its patience being 1 epoch. With this decaying scheme, we pre-trained the model for 10 epochs with an initial learning rate of 0.0003 and fine-tuned the model for 10 epochs with an initial learning rate of 0.00005.

D. Ablation Studies

To provide a deeper understanding of our approach, we conducted ablation studies to investigate the influence of several components in our predictive system.

Ablation Study 1: In this ablation study, we aimed at evaluating the influence of using non-glucose features in our features set. To this end, we designed the experiment of reducing the features set to only glucose values, where the problem correspondingly reduced to a uni-modal time-series prediction problem.

Ablation Study 2: We also evaluated the effects of using transfer learning, i.e., the effects of using data from other subjects to pre-train the model. With this aim, we replaced the data of other subjects in our pre-training phase with the target subject's personal training data. This led to a two-stage training process on a subject's own data, where each stage was trained using a different initial learning rate.

E. Evaluation Metrics

Following previous studies, we used Root Mean Square Error (RMSE) as the evaluation metric, calculated by

$$\text{RMSE} = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(\hat{g}_j - g_j)^2} \tag{7}$$

where $\hat{g}$ and g were the prediction and ground truth of glucose values in mg/dL, respectively, and G was the total number of predictions in the test set.

To evaluate the statistical significance of our results, we employed the one-sided paired T-test with $p = 0.05$ as threshold. When comparing our approach to the baselines, we used subject-level RMSEs as the two groups of samples in the T-tests, and we tested if our results are statistically significantly lower than the best available baseline. When comparing our approach to the ablation experiments, we used the squared errors of all predictions on the test set as the two groups of samples in the T-tests, and we tested if our results are statistically significantly lower than the counterpart results in the ablation experiments.

V. RESULTS

Main experiments: Comparing to recent methods for prediction horizons of 30 minutes and 60 minutes (Tables I and

[1]Only a subset of results on the full 12 subjects may be present in the baselines since the two OhioT1DM subsets were released separately with a two-year interval. Additionally, the subject-level results in the baseline [33] were not provided by the authors.

[2]A bold value indicates that it is the lowest in its comparisons.

[3]A result marked by an asterisk * is statistically significantly ($p < 0.05$) lower than its best available baseline.

TABLE III
COMPARING ABLATION STUDY 1 WITH THE MAIN EXPERIMENTS: EFFECTS OF USING GLUCOSE-RELATED EVENTS IN THE FEATURES SET.[2,4]

Subject	Feature		RMSE (mg/dL) *540*	*544*	*552*	*567*	*584*	*596*	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
30 min	Glucose	-	21.95	17.66	16.67	20.53	21.47	17.02	18.47	**18.06**	15.29	**21.31**	17.46	20.98	18.90 (2.23)
	Full	Set. 1	**19.49***	15.75*	15.68*	**18.96***	**19.52***	15.94*	**17.57**	18.34	14.84	21.69	16.02*	20.08	**17.82 (2.17)**
		Set. 2	20.24*	**15.26***	**15.49***	19.72	20.23*	**15.90***	17.59	18.34	**14.64**	**21.31**	**15.93***	**20.03**	17.89 (2.37)
60 min	Glucose	-	38.96	29.81	29.40	36.14	35.01	28.07	31.59	28.99	25.98	32.93	29.02	32.32	31.52 (3.74)
	Full	Set. 1	**32.79***	**25.23***	**25.71***	**30.87***	**31.20***	**25.14***	29.56*	**28.37**	**25.04**	**32.34**	**26.73***	**29.56***	**28.54 (2.91)***
		Set. 2	38.74	26.57*	28.56	34.49	31.48*	26.23*	**29.31***	29.82	26.22	32.61	27.61	31.66	30.27 (3.77)

TABLE IV
COMPARING ABLATION STUDY 2 WITH THE MAIN EXPERIMENTS: EFFECTS OF USING DATA FROM OTHER SUBJECTS.[2,4]

Subject	Data		RMSE (mg/dL) *540*	*544*	*552*	*567*	*584*	*596*	*559*	*563*	*570*	*575*	*588*	*591*	Avg (SD)
30 min	Personal	Set. 1	19.97	15.69	16.12	19.05	20.68	16.81	18.93	**18.07**	15.66	21.85	16.60	20.68	18.34 (2.16)
		Set. 2	21.63	16.03	17.00	19.78	20.32	16.79	18.54	18.37	15.50	22.48	16.40	20.53	18.62 (2.32)
	All	Set. 1	**19.49**	15.75	15.68	**18.96**	**19.52***	15.94*	**17.57***	18.34	14.84	21.69	16.02	20.08	**17.82 (2.17)**
		Set. 2	20.24*	**15.26***	**15.49***	19.72	20.23	**15.90***	17.59	18.34	**14.64***	**21.31**	**15.93**	**20.03**	17.89 (2.37)
60 min	Personal	Set. 1	34.68	26.03	26.16	31.44	31.94	27.80	29.53	**27.98**	26.45	34.07	27.56	31.07	29.56 (3.03)
		Set. 2	37.79	26.06	28.42	34.08	32.47	27.42	30.60	30.39	26.36	33.82	28.39	31.49	30.61 (3.53)
	All	Set. 1	**32.79***	**25.23**	**25.71**	**30.87**	**31.20**	**25.14***	29.56	28.37	**25.04***	**32.34**	**26.73**	**29.56***	**28.54 (2.91)***
		Set. 2	38.74	26.57	28.56	34.49	31.48	26.23	**29.31**	29.82	26.22	32.61	27.61	31.66	30.27 (3.77)

II), our approach achieved state-of-the-art performance with a statistically significant reduction of RMSE in both setting 1 and 2. Moreover, since glucose-related events (changing basal rate, taking extra insulin bolus and having food) take time to affect the glucose concentration, the difference of predictive performance between setting 1 and 2 was more noticeable in 60 minutes prediction horizon than 30 minutes.

Ablation studies: Table III shows that the presence of insulin and carbohydrate intake information increased the performance of our approach on most subjects. This observation demonstrated that our model was able to learn the effects of insulin and carbohydrates on glucose and make use of them during inference. Additionally, as shown in Table IV, results were better when transfer learning was applied, indicating that data from other people with T1DM helped the personalised learning.

VI. DISCUSSION

In the main experiments, our approach had better performance in setting 1 than setting 2 ($p < 0.05$) in the 60 minutes prediction horizon, while a similar result was not observed in 30 minutes prediction horizon. This observation was consistent with [33], indicating that knowing the glucose-related events was more helpful in longer-term glucose prediction. Therefore, we argue that the prediction could be more accurate if the ongoing glucose-related factors are continuously provided to adjust the prediction. Besides, the better results of applying transfer learning than using only personal data in ablation study 2 suggested that (i) larger amounts of personal data can further improve the predictive performance, and (ii) data from other people could serve as data augmentation when personal data are insufficient for training a glucose predictive system.

[4]A result marked by an asterisk * is statistically significantly ($p < 0.05$) lower than its counterpart result in the ablation experiments.

Regarding the limitation, we did not fully experiment on the other features in the OhioT1DM dataset such as heart rate and galvanic skin response, because in our pilot experiments they showed marginal improvement of the glucose prediction performance. However, further investigation is needed to fully explore the potential of these features in glucose prediction as well as to identify the optimal way to incorporate them into our proposed model architecture. Another future work is research on the self-adjustable predictive algorithm through time which suits real-life applications better.

VII. CONCLUSION

In this paper, we have introduced an approach for personalised short-term glucose prediction for people with T1DM based on a combined short history of glucose values and glucose-related events. Our approach achieves state-of-the-art results compared to recent glucose prediction baselines. Experiments suggest that eight weeks of personal data are still inadequate for deep learning methods to achieve their best potential, and that when the personal data are insufficient, one could use data from other people with T1DM towards helping the model to achieve better performance. Moreover, our experiments show that announcing the planned glucose-related events such as food intake and insulin to be taken in near future is helpful for predicting glucose, especially in longer prediction horizons.

ACKNOWLEDGEMENTS

This research has been delivered in partnership with and funded by Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. We gratefully acknowledge the funding from the ANU School of Computing for the first author's PhD studies,

and the Ohio University as the source of the OhioT1DM dataset. We would also like to thank the GPU clusters provided by National Computational Infrastructure Australia.

REFERENCES

[1] C. Dalla Man, R. A. Rizza, and C. Cobelli, "Meal simulation model of the glucose-insulin system," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 10, pp. 1740–1749, 2007.

[2] R. Hovorka, V. Canonico, L. J. Chassin, U. Haueter, M. Massi-Benedetti, M. O. Federici, T. R. Pieber, H. C. Schaller, L. Schaupp, T. Vering *et al.*, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiological Measurement*, vol. 25, no. 4, p. 905, 2004.

[3] N. Brew-Sam, M. Chhabra, A. Parkinson, K. Hannan, E. Brown, L. Pedley, K. Brown, K. Wright, E. Pedley, C. J. Nolan *et al.*, "Experiences of young people and their caregivers of using technology to manage type 1 diabetes mellitus: Systematic literature review and narrative synthesis," *Journal of Medical Internet Research Diabetes*, vol. 6, no. 1, p. e20973, 2021.

[4] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, "Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network," in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*. IEEE, 2018, pp. 1–5.

[5] D.-Y. Kim, D.-S. Choi, J. Kim, S. W. Chun, H.-W. Gil, N.-J. Cho, A. R. Kang, and J. Woo, "Developing an individual glucose prediction model using recurrent neural network," *Sensors*, vol. 20, no. 22, p. 6460, 2020.

[6] F. Allam, Z. Nossai, H. Gomma, I. Ibrahim, and M. Abdelsalam, "A recurrent neural network approach for predicting glucose concentration in type-1 diabetic patients," in *Engineering Applications of Neural Networks*. Springer, 2011, pp. 254–259.

[7] J. Martinsson, A. Schliep, B. Eliasson, C. Meijner, S. Persson, and O. Mogren, "Automatic blood glucose prediction with confidence using recurrent neural networks," in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data*. CEUR Workshop Proceedings, 2018, pp. 64–68.

[8] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 28, no. 3. PMLR, 17–19 Jun 2013, pp. 1310–1318.

[9] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *IEEE International Conference on Neural Networks*. IEEE, 1993, pp. 1183–1188.

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

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[12] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1459–1469.

[13] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," *arXiv preprint arXiv:2010.05468*, 2020.

[14] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring cross-domain knowledge for video sign language recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6205–6214.

[15] D. Li, C. Xu, K. Zhang, X. Yu, Y. Zhong, W. Ren, H. Suominen, and H. Li, "Arvo: Learning all-range volumetric correspondence for video deblurring," *arXiv preprint arXiv:2103.04260*, 2021.

[16] A. Bock, G. François, and D. Gillet, "A therapy parameter-based model for predicting blood glucose concentrations in patients with type 1 diabetes," *Computer Methods and Programs in Biomedicine*, vol. 118, no. 2, pp. 107–123, 2015.

[17] C.-L. Chen and H.-W. Tsai, "Modeling the physiological glucose–insulin system on normal and diabetic subjects," *Computer Methods and Programs in Biomedicine*, vol. 97, no. 2, pp. 130–140, 2010.

[18] Z. Wu, C.-K. Chui, G.-S. Hong, and S. Chang, "Physiological analysis on oscillatory behavior of glucose–insulin regulation by model with delays," *Journal of Theoretical Biology*, vol. 280, no. 1, pp. 1–9, 2011.

[19] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[20] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, "A review of personalized blood glucose prediction strategies for T1DM patients," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 33, no. 6, p. e2833, 2017.

[21] T. Bremer and D. A. Gough, "Is blood glucose predictable from previous values? a solicitation for data." *Diabetes*, vol. 48, no. 3, pp. 445–451, 1999.

[22] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 5, pp. 931–937, 2007.

[23] A. Gani, A. V. Gribok, Y. Lu, W. K. Ward, R. A. Vigersky, and J. Reifman, "Universal glucose models for predicting subcutaneous glucose concentration in humans," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 157–165, 2009.

[24] V. Naumova, S. V. Pereverzyev, and S. Sivananthan, "A meta-learning approach to the regularized learning—case study: Blood glucose prediction," *Neural Networks*, vol. 33, pp. 181–193, 2012.

[25] S. Shanthi, P. Balamurugan, and D. Kumar, "Performance comparison of featured neural network with gradient descent and Levenberg-Marquart algorithm trained neural networks for prediction of blood glucose values with continuous glucose monitoring sensor data," in *2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET)*. IEEE, 2012, pp. 385–391.

[26] C. Pérez-Gandía, A. Facchinetti, G. Sparacino, C. Cobelli, E. Gómez, M. Rigla, A. de Leiva, and M. Hernando, "Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring," *Diabetes Technology & Therapeutics*, vol. 12, no. 1, pp. 81–88, 2010.

[27] S. Fong, Y. Zhang, J. Fiaidhi, O. Mohammed, and S. Mohammed, "Evaluation of stream mining classifiers for real-time clinical decision support system: a case study of blood glucose prediction in diabetes therapy," *BioMed Research International*, vol. 2013, pp. 1–16, August 2013.

[28] C. Marling and R. C. Bunescu, "The OhioT1DM dataset for blood glucose level prediction," in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data*. CEUR Workshop Proceedings, 2018, pp. 60–63.

[29] K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 603–613, 2019.

[30] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction." in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data*. CEUR Workshop Proceedings, 2018, pp. 64–78.

[31] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *Journal of Healthcare Informatics Research*, vol. 4, no. 3, pp. 308–324, 2020.

[32] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *5th International Workshop on Knowledge Discovery in Healthcare Data*. CEUR Workshop Proceedings, 2020.

[33] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "LSTMs and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data," in *The 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 706–712.

[34] T. Zhu, X. Yao, K. Li, P. Herrero, and P. Georgiou, "Blood glucose prediction for type 1 diabetes using generative adversarial networks," in *The 5th International Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675. CEUR Workshop Proceedings, 2020, pp. 90–94.

[35] A. Bertachi, L. Biagi, I. Contreras, N. Luo, and J. Vehí, "Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks." in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data*. CEUR Workshop Proceedings, 2018, pp. 85–90.

[36] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Data fusion of activity and CGM for predicting blood glucose levels," in *The 5th International Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675. CEUR Workshop Proceedings, 2020, pp. 120–124.

Meta-optimised Time-index Modeling for Glucose Excursion Prediction

Ran Cui[a], Hanna Suominen[a,b,c], Christopher J Nolan[c], and Elena Daskalaki[a]
[a] School of Computing, The Australian National University, Canberra, Australia
[b]Department of Computing, University of Turku, Turku, Finland
[c]School of Medicine and Psychology, The Australian National University, Canberra, Australia
Emails: {ran.cui, hanna.suominen, christopher.nolan, eleni.daskalaki}@anu.edu.au

***Abstract*—The development of continuous glucose monitoring (CGM) devices has facilitated research of data-driven glucose prediction, which is useful in managing type 1 diabetes mellitus (T1DM). Existing research has investigated using historical-value based machine learning approaches to predict short-term glucose excursion and adapting pre-trained models to unseen patients via meta-learning. This study aims to investigate if time-index based modeling could improve the effectiveness and efficiency to the same objective. Different to existing methods that take the past glucose sequence as the input and output the future sequence, the method we apply — referred to as the meta-optimised time-index model (MOTIM) — takes time-index features as the input and predicts the glucose value at that time. We argue that the reduction of modeling complexity from sequence-level to value-level could ease the learning difficulty and increase the efficiency. In our experiments on the OhioT1DM and UMT1DM datasets, the MOTIM achieves comparable performance to the state-of-the-art models (e.g., 1-hour excursion averaged mean absolute percentage error of 10.25 on OhioT1DM and 12.50 on UMT1DM) while consuming a substantially lower cost (e.g., only 1.3% model parameters and more than 10 times faster inference). These promising findings encourage further work on time-index modeling in glucose prediction for T1DM, and to encourage it, we have released our codes at `https://github.com/r-cui/MOTIMGluPred` under the MIT license.**

***Index Terms*—glucose prediction, time-index model, machine learning, type 1 diabetes mellitus, evaluation study**

I. Introduction

The development of continuous glucose monitoring (CGM) devices [1] has enabled people with type 1 diabetes mellitus (T1DM) to constantly track their glucose trajectory in order to support disease management. Data-driven personalised glucose prediction using machine learning techniques, as a research topic derived from this technical innovation, has also drawn extensive attention [2].

To date, existing research on data-driven glucose prediction has investigated using historical-value models, as illustrated in Figure 1(a), to predict a future glucose value, such as $\hat{g}_{t+1}$. To achieve this, they learn a mapping from the most recent glucose sequence of the form

$$\hat{g}_{t+1} = f(g_t, g_{t-1}, \ldots, g_{t-I+1})$$

This research has been delivered in partnership with and funded by Our Health in Our Hands (OHIOH), a strategic initiative of The Australian National University (ANU). We gratefully acknowledge the funding from the ANU School of Computing for the first author's PhD studies.

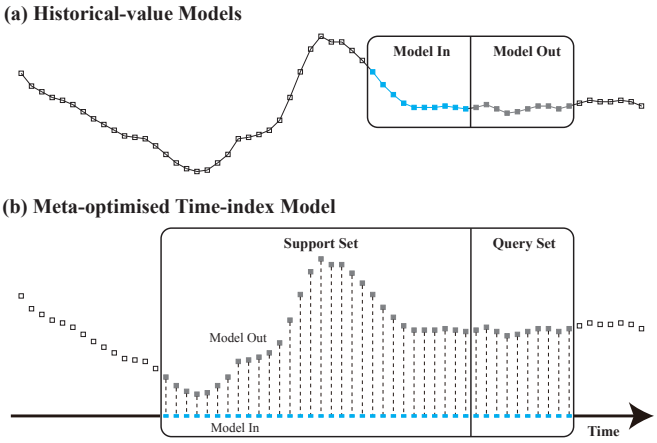


Fig. 1. Illustration of the key difference between historical-value models and the MOTIM in predicting glucose excursion. We mark a larger lookback window in the MOTIM to make distinction to the shorter input length in the historical-value models literature. Figure is adapted from [3].

where t denotes the current time index and I denotes the length of the input glucose sequence. As reviewed in [2], historical-value model based glucose prediction has yielded promising outcomes and has become a methodological paradigm for data-driven glucose prediction. Following this, a line of research has further investigated cross-instance adaptation, which considers non-personalised learning scanario and targets using data from other patients to aid the learning of a personalised model [4]–[9]. This could be useful in particular during the early stage of a patient's CGM-device use, as the amount of personalised data is limited. However, regardless of the task being personalised learning or cross-instance adaptation, historical-value models need to properly learn the distribution of the target glucose values conditioned on the input sequence in $\mathbb{R}^I$. We make the argument that this high dimensional sequence-level conditional distribution could cause an intrinsic learning difficulty, especially in the scenario of cross-instance adaptation where the input sequence distribution $p(g_t, g_{t-1}, \ldots, g_{t-I+1})$ would naturally shift.

In this study, we raise the research question of how the aforementioned difficulty can be mitigated by improving the modeling. Inspired by the recent advancement in deep learning for generic time-series forecasting and the success of meta-

learning in cross-instance adaptation, we apply the meta-optimised time-index model (MOTIM) [3] to data-driven glucose prediction. Different to historical-value models, the MOTIM employs a deep time-index model [10], which takes a time-index feature as its input and outputs the corresponding glucose value at that time-index, as illustrated by the vertical connections in Figure 1(b). More formally, it learns the following mapping:

$$\hat{g}_t = f(\boldsymbol{\tau}_t),$$

where $\boldsymbol{\tau}_t$ is the time-index feature of a specific time-index t. In contrast to learning a distribution $p(g_{t+1}|g_t, g_{t-1}, \ldots, g_{t-I+1})$ as in historical-value models, the MOTIM learns $p(g_t|\boldsymbol{\tau}_t)$, thus the learning complexity is reduced from sequence-level to value-level. The cost of this simplification is that a sole deep time-index model is limited in making prediction using future time-indices. Thus, the MOTIM uses its meta-optimisation framework to tune the time-index model closer to the local inductive bias before the extrapolation. In particular, for predicting an excursion as marked as the query set in Figure 1(b), the last layer of the MOTIM is optimised by referring to the examples in the lookback window while the rest of the parameters are optimised globally using all available data. To this end, the MOTIM retains the capacity of predicting the future sequence by examining the past sequence, akin to historical-value models, while the modeling is purely time-index based which can decouple the learning of sequence-level mapping as in historical-value models.

Apart from serving as a novel substitution of the traditional historical-value models under the standard personalised training paradigm, the MOTIM is also synergistic to cross-instance adaptation in glucose prediction. As an analogy, we could understand the multi-patient training data as a single concatenated long glucose sequence where the glucose distribution changes at the points where the source patient changes. The meta-optimisation would then fit the model to the local bias, that is, the personalised pattern, in this context. To validate the effectiveness, we evaluate the MOTIM in comparison to existing historical-value models in both the standard personalised training, and cross-instance adaptation using pre-training & fine-tuning. Our experiments on the OhioT1DM [11] and UMT1DM [12] datasets indicate that the MOTIM achieves predictive performance comparable to the state of the art. Additionally, the MOTIM gives significantly higher efficiency than the historical-value models both in memory requirement and time efficiency. Last but not least, we adopt glucose excursion prediction [12]–[14] as the prediction target, which requires predicting the full glucose trajectory in the prediction horizon, instead of a single-point prediction.

We summarise our contribution in this study as follows:

- We propose for the first time to apply the MOTIM, which is novel to existing historical-value models to data-driven glucose excursion prediction. The method achieves predictive performance comparable to the state of the art.
- We demonstrate that the MOTIM also achieves state-of-the-art predictive performance in cross-instance adaptation of glucose excursion prediction, where pre-training & fine-tuning method is applied.
- We show the significantly higher efficiency of the MOTIM compared to existing historical-value based glucose prediction models which would facilitate the implementation in real life on smart devices.

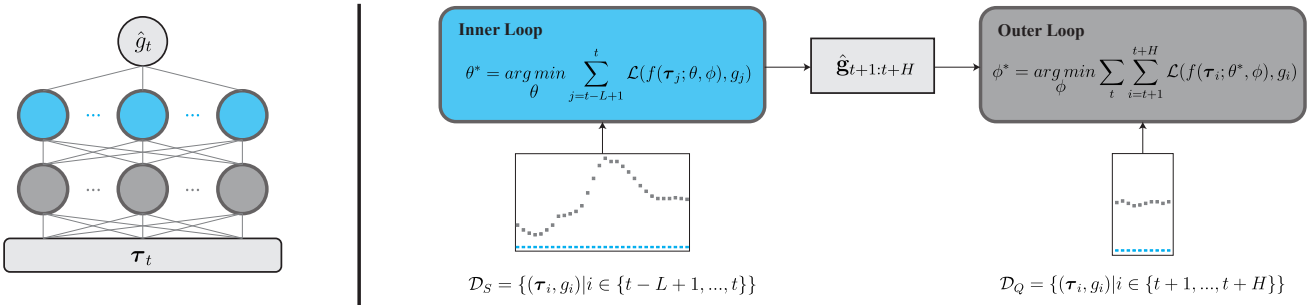
II. Related Works

A. Glucose Excursion Prediction

Glucose excursion prediction, or multi-step glucose prediction [14], is a specific task in the field of glucose prediction [2] which requires predicting the full trajectory in the prediction horizon instead of at a single point. Despite the increased difficulty, patients can make better decisions based on a reliable excursion prediction than merely the prediction of one point. A common ground of existing data-driven methods on this task is that the model takes the past sequence as its input, hence we term this approach as historical-value modeling. For example, [15] adopts a temporal convolutional network (TCN) [16] as the backbone and uses the sequence-to-sequence [17] scheme to recursively decode the latent state encoded from the input sequence to predict the excursion. [4] employs the self-attention mechanism [18] as the encoder and applies the same model recursively to produce the values in the prediction horizon one at a time. To circumvent the error propagation problem in the recursive output scheme of the aforementioned studies, [12] proposes to apply the multi-output strategy [19] at the output end which links the same gated recurrent unit (GRU) [20] to multiple output heads to predict different values in the prediction horizon. However, as the prediction of these methods are obtained via a direct learned mapping from the recently observed glucose sequence, learning the high-dimensional conditional distribution remains a difficulty which we aim to improve upon.

B. Cross-instance Adaptation in Glucose Prediction

Cross-instance adaptation is another research topic in data-driven glucose prediction, which targets the difficulty in obtaining sufficient personalised data to train personalised models. The core research question is how to let other patients' data contribute to the learning of a personalised glucose prediction model. The most explored methods for this purpose are the transfer-learning [4]–[7] and meta-learning [8], [9]. In these methods, data from other patients are first used to train the deep model so that the model learns a general glucose pattern of people with T1DM. The generic model is then fine-tuned to an unseen patient by further training using personalised data. Another line of research has focused on designing multi-task models to enable learning towards a specific subset of patients. For example, [21] divides the patients into two groups based on gender and connects a shared long short-term memory (LSTM) [22] backbone to two different fully connected output layers to enable a gender-specific learning. [23] further assigns patient-specific fully connected heads to all subjects on top



(a) Time-index model. (b) Meta-optimisation framework.

Fig. 2. (Left) The time-index model architecture, which contains K-layered implicit neural representations (INR) (grey nodes) to encode the deep time features and ridge regressor (blue nodes) to generate the predicted glucose values. (Right) The meta-optimisation framework for the time-index model. We make use of blue and grey nodes to mark the correspondences of the group of parameters in the model to be optimised in different stages of the meta-learning. Figure is adapted from [3].

of a shared LSTM backbone. Our applied method called the MOTIM relates to these studies as it incorporates a meta-optimisation framework with the deep time-index model, thus naturally fits to the task of cross-instance adaptation.

III. Methodology

The aim of glucose excursion prediction is to predict the full trajectory of glucose readings within a prediction horizon. By indexing the current time as t, the target is to obtain a predictive model $f(\cdot)$, which predicts the glucose sequence $\mathbf{g}_{t+1:t+H} = [g_{t+1}, g_{t+2}, \ldots, g_{t+H}]$.

A. Historical-value Models

We precede our description of the MOTIM method by a formal formulation of historical-value models, so distinction can be made. The central idea of historical-value models is that prediction of $\mathbf{g}_{t+1:t+H}$ is obtained via a mapping from a recent glucose sequence $\mathbf{g}_{t-I+1:t} = [g_{t-I+1}, g_{t-I+2}, \ldots, g_t]$. More formally, historical-value models learn a mapping of the form

$$f(\mathbf{g}_{t-I+1:t}) = \hat{\mathbf{g}}_{t+1:t+H}. \tag{1}$$

For example, a basic linear model learns a linear mapping $\mathbb{R}^I \rightarrow \mathbb{R}^H$. In more complex deep models, the input sequence is usually first encoded into a deep feature in $\mathbb{R}^D$ where D denotes a latent dimension, and the $\mathbb{R}^D$ feature is then projected to the final output in $\mathbb{R}^H$.

Therefore, historical-value models learn the conditional distribution

$$p(g_{t+1}, g_{t+2}, \ldots, g_{t+H} | g_t, g_{t-1}, \ldots, g_{t-I+1}).$$

The complex nature of this distribution, that a sequence of future glucose values are conditioned on a sequence of observed glucose values, creates an intrinsic difficulty in the learning. For the same reason, existing methods in this category usually learn from relatively short past glucose sequences, such as from 1 to 2 hours (from 6 to 12 data points) [4], [12], [15]. In cross-instance adaptation, it is probable that the new input sequence $\mathbf{g}_{t-I+1:t}$ under a shifted distribution has not been adequately learned [24]. To tackle this, existing studies on glucose prediction using historical-value models have investigated the pipeline of fine-tuning a pre-trained model to unseen patients targeting the glucose pattern distribution shift among different patients [4]–[9]. Among such studies, several of them have chosen to formalise the pipeline using a meta-learning framework (Table I).

B. Meta-optimised Time-index Model

Different to historical-value models, the MOTIM [3] employs a deep time-index model [10] supported by a meta-optimisation framework. Unlike historical-value models, the learning objective of a time-index model is a mapping from a time-index feature to the glucose value at that time-index, shown by the vertical connection in Figure 1(b). Although a deep time-index model can effectively fit the observed time steps, it has been shown to be limited in predicting the future via merely the extrapolation of time-index features [3]. To tackle this, the model is enabled to also learn inductive bias brought by the most recent glucose pattern within a lookback window by the meta-optimisation framework.

1) Deep Time-index Model: The deep time-index model in the MOTIM learns the following mapping:

$$f(\boldsymbol{\tau}_t) = \hat{g}_t.$$

It takes the Gaussian Fourier features augmented relative coordinates as the time-index feature [25]. More specifically, for a one-dimensional coordinates vector $\mathbf{c}$ of length $L + H$ where L denotes the length of the lookback window, the high-dimensional time-index feature $\boldsymbol{\tau}$ is given by the transformation $\gamma(\cdot)$ of form

$$\gamma(\mathbf{c}) = [\sin(2\pi\mathbf{B}\mathbf{c}), \cos(2\pi\mathbf{B}\mathbf{c})] \in \mathbb{R}^{(L+H)\times D}.$$

Here, D denotes the hidden dimension of the model, $\mathbf{B} \in \mathbb{R}^{D/2}$ is a non-trainable parameter sampled from Gaussian distribution $\mathcal{N}(0, \sigma^2)$ scaled with hyperparameter σ^2, and the coordinates vector $\mathbf{c}$ is a $[0, 1]$ normalised time-indices vector given by

$$\mathbf{c}_{t+i} = \frac{i + L}{L + H - 1}$$

TABLE I
CONCEPTUAL COMPARISON BETWEEN META-LEARNING WITH HISTORICAL-VALUE MODELS AND THE MOTIM.

	Historical-value Model	MOTIM
Modeling	$f(\mathbf{g}_{t-I+1:t}) = \mathbf{g}_{t+1:t+H}$	$f(\boldsymbol{\tau}_t) = \hat{g}_t$
Task	existing patients $\rightarrow$ new patient	$\mathbf{g}_{t-L+1:t} \rightarrow \mathbf{g}_{t+1:t+H}$
Support Set	existing patients	$\{(\boldsymbol{\tau}_i, g_i) \vert i \in \{t-L+1, \dots, t\}\}$
Query Set	new patient	$\{(\boldsymbol{\tau}_i, g_i) \vert i \in \{t+1, \dots, t+H\}\}$

for $i = -L, \dots, H-1$. To this end, at each position in the window $L+H$, its time feature becomes a concatenation of $D/2$ Fourier features with different frequencies, thus ensures a sufficiently expressive yet distinguishable representation across all time-indices.

As illustrated in Figure 2(a), the time-index feature $\boldsymbol{\tau}_t$ is then encoded by the implicit neural representations (INR) module [26], which is a K-layered, rectified linear unit (ReLU) activated fully connected network, to learn the high dimensional time features. This encoding is given by

$$\begin{aligned} \mathbf{z}^{(0)} &= \boldsymbol{\tau}, \\ \mathbf{z}^{(k+1)} &= \max(0, \mathbf{z}^{(k)}\mathbf{W}^{(k)} + \mathbf{b}^{(k)}), \end{aligned}$$

in which $\mathbf{W}^{(k)} \in \mathbb{R}^{D\times D}$, $\mathbf{b}^{(k)} \in \mathbb{R}^{D}$, and $\mathbf{z}^{(k)} \in \mathbb{R}^{(L+H)\times D}$ for k in $0, \dots, K-1$ constitute the trainable parameters of the INR. The final output of the INR, $\mathbf{z}^{(K)}$, is the learned position-wise representation. To produce the predicted glucose values, a final ridge regressor layer [27] is subsequently applied, given by

$$\hat{\mathbf{g}} = \mathbf{z}^{(K)}\mathbf{w}_R + b_R, \tag{2}$$

where $\mathbf{w}_R \in \mathbb{R}^D$ and $b_R \in \mathbb{R}$ are the trainable parameters. This linear connection based network structure provides architectural assurance for the efficiency of the MOTIM.

2) Meta-optimisation Framework: The meta-optimisation framework of the MOTIM enables the time-index model to refer to the recent glucose pattern when making an excursion prediction, otherwise the prediction is prone to be weak [3]. Note that the meta-optimisation framework in the MOTIM is part of the method itself, meaning the meta-optimisation is performed regardless of the prediction scenario being cross-instance or not. To this end, it is considered more general than historical-value models using meta-learning, as the meta-learning for the latter only exists in cross-instance scenario. At each time t where a future excursion prediction of length H is desired, the observed examples in the lookback window

$$\mathcal{D}_S = \{(\boldsymbol{\tau}_i, g_i) \vert i \in \{t-L+1, \dots, t\}\}$$

are treated as the support set, and the examples in the prediction horizon

$$\mathcal{D}_Q = \{(\boldsymbol{\tau}_i, g_i) \vert i \in \{t+1, \dots, t+H\}\}$$

serve as the query set. A comparison between the meta-learning concepts in historical-value models and in the MOTIM is summarised in Table I.

To achieve this, the parameters of the full model are divided into two groups

$$\begin{aligned} \theta &= \{\mathbf{w}_R, b_R\} \text{ and} \\ \phi &= \{\mathbf{W}^{(k)}, \mathbf{b}^{(k)} \vert k \in \{0, \dots, K-1\}\} \end{aligned}$$

to be optimised in different phases, as distinguished by different colors in Figure 2(a). The formulated optimisation target of this bi-level optimisation problem is given by

$$\phi^* = \arg\min_{\phi} \sum_{t} \sum_{i=t+1}^{t+H} \mathcal{L}(f(\boldsymbol{\tau}_i; \theta^*, \phi), g_i) \tag{3}$$

$$\text{s.t. } \theta^* = \arg\min_{\theta} \sum_{j=t-L+1}^{t} \mathcal{L}(f(\boldsymbol{\tau}_j; \theta, \phi), g_j), \tag{4}$$

where $\mathcal{L}(\cdot,\cdot)$ denotes the loss function, meaning that the outer loop (Equation (3)) optimises ϕ over all possible windows of length $(L+H)$ from the training data in which in each instance the parameter θ has been optimised by the inner loop (Equation (4)). In other words, the training is essentially the learning of the INR representation $\mathbf{z}^{(K)}$ via the outer loop in Equation (3), while the fit to local glucose pattern within the lookback window is carried out by the inner loop in Equation (4).

The inner loop optimisation targets only the optimisation of the ridge regressor (Equation (2)), i.e., find $\mathbf{w}_R^*$ and b_R^* that best fit $\mathbf{g} = \mathbf{z}^{(K)}\mathbf{w}_R + b_R$. Note that this optimisation involves only the inner loop, both the glucose values $\mathbf{g}$ and the INR output $\mathbf{z}^{(K)}$ are given. Therefore, close-form solution $\mathbf{w}^* = [\mathbf{w}_R^*, b_R^*]$ can be obtained by analytically solving the linear system, given by

$$\begin{aligned} \mathbf{w}^* &= \arg\min_{\mathbf{w}} \|\mathbf{z}^{(K)}\mathbf{w} - \mathbf{g}\|^2 + \lambda\|\mathbf{w}\|^2 \\ &= (\mathbf{z}^{(K)^T}\mathbf{z}^{(K)} + \lambda\mathbf{I})^{-1}\mathbf{z}^{(K)^T}\mathbf{g} \end{aligned}$$

where λ is the regularisation coefficient as depicted in [27]. This differentiable close-form solution of inner loop enables the outer loop to be optimised by regular gradient-based back-propagation learning [28]. An illustration of this meta-optimisation process is shown in Figure 2(b).

IV. EXPERIMENTS

A. Experiment Setups

We conduct the following experiments to validate the effectiveness of the MOTIM. In all experiments, we report the `mean(std)` of the same 12 randomly selected patients under a 60-minute prediction horizon.

1) Personalised Training Experiment: Each model is both trained and evaluated using the full training and test data from the same diabetes patient in this experiment. Thus, it reflects the standard personalised application of glucose prediction.

TABLE II
PERSONALISED TRAINING EXPERIMENT RESULTS. SCORES WITH BEST MEAN VALUE ARE MARKED AS BOLD, AND THE MOTIM RESULTS THAT ARE SIGNIFICANTLY BETTER ($p < 0.05$) THAN AT LEAST HALF OF THE BENCHMARKS ARE UNDERLINED.

	Model	RMSE↓			MAPE↓			TP ACC↑		Abnormal glycemia MCC↑	
		all	at 30 min	at 60 min	all	at 30 min	at 60 min	δ=1	δ=2	Hyperglycemia	Hypoglycemia
OhioT1DM	*Repeat*	25.73 (2.98)	23.28 (2.74)	38.16 (4.45)	11.64 (2.03)	11.35 (1.98)	19.29 (3.44)	-	-	74.83 (5.10)	43.67 (19.61)
	Linear	24.38 (2.61)	22.21 (2.75)	35.62 (3.87)	11.54 (2.06)	11.13 (2.05)	18.89 (3.63)	57.82 (2.11)	59.67 (2.10)	75.17 (4.85)	37.06 (13.26)
	MoGRU	22.64 (2.45)	**20.17 (2.42)**	33.78 (3.55)	10.55 (2.12)	10.08 (2.07)	17.56 (3.32)	**60.78 (2.59)**	**61.95 (2.68)**	**78.45 (4.45)**	33.90 (17.41)
	ReSelfAttn	24.87 (3.14)	22.63 (3.05)	35.76 (4.15)	11.99 (2.58)	11.45 (2.53)	18.71 (3.85)	60.31 (3.26)	61.31 (3.29)	75.41 (4.67)	24.85 (17.64)
	SeqConv	25.95 (2.35)	23.65 (2.27)	37.25 (3.33)	13.10 (2.35)	12.50 (2.21)	20.54 (3.75)	59.71 (2.91)	60.30 (2.97)	73.08 (6.30)	16.74 (13.19)
	MOTIM	**22.44 (2.75)**	20.24 (2.83)	**33.17 (3.80)**	**10.25 (1.81)**	**9.86 (1.76)**	**17.28 (3.23)**	60.06 (2.85)	61.05 (2.97)	78.19 (4.10)	**41.54 (15.35)**
UMT1DM	*Repeat*	28.13 (6.07)	25.93 (5.64)	40.64 (9.19)	12.79 (4.03)	12.58 (4.01)	20.30 (6.45)	-	-	72.73 (9.30)	50.26 (25.43)
	Linear	26.41 (5.72)	24.79 (5.71)	37.20 (8.34)	13.64 (4.76)	13.33 (4.64)	21.65 (7.76)	53.89 (3.30)	56.56 (3.20)	73.93 (10.35)	**44.76 (18.18)**
	MoGRU	25.68 (5.89)	23.77 (5.71)	36.55 (8.36)	12.86 (4.67)	12.43 (4.69)	20.89 (7.76)	55.00 (4.16)	56.73 (3.45)	**74.22 (9.64)**	43.41 (23.19)
	ReSelfAttn	27.57 (5.76)	25.57 (5.48)	38.46 (8.44)	14.36 (5.87)	13.84 (5.76)	21.24 (7.88)	**56.25 (3.73)**	57.48 (3.73)	73.20 (10.26)	39.47 (25.13)
	SeqConv	28.79 (6.66)	26.73 (6.17)	40.27 (9.68)	15.97 (6.73)	15.40 (6.52)	24.17 (9.70)	54.38 (4.59)	55.46 (4.10)	70.85 (8.53)	28.22 (16.96)
	MOTIM	**25.31 (5.68)**	**23.62 (5.78)**	**35.75 (8.06)**	**12.50 (4.71)**	**12.09 (4.68)**	**20.23 (7.92)**	54.82 (4.45)	**56.90 (4.20)**	73.39 (15.93)	44.14 (25.49)

2) Cross-instance Adaptation Experiment: In this experiment, one patient is fixed as an anchor and the next patient in the patient list serves as its auxiliary patient. We evaluate first using all training data from the auxiliary patient to pre-train the model to its convergence, then fine-tuning the model using the anchor patient's personal training data. To investigate the adaptation process in details, we set a hyperparameter *portion of usage* (PoU) taking value of $\{0.0, 1.0\}$ which controls the portion of personalised training data to use in the fine-tuning. Note that in $PoU = 0.0$, the scenario reduces to removing the fine tuning stage.

B. Datasets

We employ the OhioT1DM [11] and UMT1DM [12] datasets in our experiments[1]. The CGM devices used in both datasets are with the same sampling period of 5 minutes.

1) OhioT1DM: The dataset created by the US Ohio University, called OhioT1DM [11], contains eight weeks of data from CGM devices and self-reported insulin treatments and meal events for 12 T1DM patients. We make use of the CGM recordings in this dataset and follow the train/test split officially provided by the authors of this dataset in our experiments.

2) UMT1DM: The dataset created by the US University of Michigan, which we refer to as UMT1DM [12], contains daily CGM data of 40 T1DM patients under a three-year protocol: patients are required to wear the CGM device for several consecutive days every three months. Following the authors [12], we use the last 3-month data chunk as the test set in our experiments. As several patients do not have records in the last three months, we choose to select 12 patients from the rest to align with the OhioT1DM experiments.

C. Evaluation Metrics

To have a comprehensive performance comparison, we evaluate the glucose excursion prediction performance using the following metrics.

[1]Ethical approval (2020/411) is obtained from the Human Research Ethics Committee of The Australian National University to use these de-identified datasets.

1) Numerical Accuracy: As per a generic time-series forecasting task, we adopt root mean square error (RMSE) and mean absolute percentage error (MAPE) to evaluate the numerical accuracy. The two metrics at a certain index $i = 1, \ldots, H$ in the prediction horizon are given by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum^{N}(g_{t+i} - \hat{g}_{t+i})^2},$$

$$\text{MAPE} = \frac{1}{N}\sum^{N}\left|\frac{g_{t+i} - \hat{g}_{t+i}}{g_{t+i}}\right|,$$

where N denotes the total number of testing examples. We report the numerical accuracy at the 30$^{\text{th}}$ minute and 60$^{\text{th}}$ minute, together with averaged scores over all positions in the predicted excursion.

2) Trend Polarity: Differing from the RMSE and MAPE, which only capture the absolute distance of a predicted value against the true value, we also evaluate the trend polarity (TP), which reflects the prediction quality in terms of the increasing or decreasing trend. To mitigate the effect of local noise due to the short sampling interval (5 minutes), we design the following smoothed trend polarity at a timestamp t in a predicted excursion which includes a tolerance factor δ to capture the overall polarity in a time range of $(2\delta+1)$ samples, given by

$$\text{TP}_t = \begin{cases} 1, & (\sum_t^{t+\delta} g_t - \sum_{t-\delta}^{t} g_t) > 0; \\ 0, & \text{otherwise.} \end{cases}$$

On this basis, the predicted polarity $\hat{\text{TP}}_t$ is then evaluated together with the ground truth TP_t following a standard classification setup, and we adopt the accuracy as the reported metric.

3) Abnormal Glycemia: Due to patients effort to avoid abnormal glycemia in their daily lives (especially hypoglycemia), the onsets of such events in our data are rare, even in some of our experiments there is no presence of abnormal glycemia at all. To this end, we do not follow the traditional clinical research in abnormal glycemia prediction that evaluates the onsets of such events using metrics sensitivity, specificity,

TABLE III
CROSS-INSTANCE ADAPTATION EXPERIMENTS RESULTS. SCORES WITH BEST MEAN VALUE ARE MARKED AS BOLD, AND THE MOTIM RESULTS THAT ARE SIGNIFICANTLY BETTER ($p < 0.05$) THAN AT LEAST HALF OF THE BENCHMARKS ARE UNDERLINED.

	PoU	Model	RMSE↓			MAPE↓			TP ACC↑		Abnormal glycemia MCC↑	
			all	at 30 min	at 60 min	all	at 30 min	at 60 min	δ=1	δ=2	Hyperglycemia	Hypoglycemia
OhioT1DM	-	*Repeat*	25.73 (2.98)	23.28 (2.74)	38.16 (4.45)	11.64 (2.03)	11.35 (1.98)	19.29 (3.44)	-	-	74.83 (5.10)	43.67 (19.61)
	0.0	Linear	24.57 (2.52)	22.67 (2.73)	35.90 (3.84)	11.75 (1.91)	11.47 (1.88)	19.33 (3.34)	57.39 (1.86)	59.09 (2.13)	75.34 (5.19)	37.45 (13.84)
		MoGRU	23.04 (2.61)	20.47 (2.63)	34.48 (3.73)	10.87 (1.88)	10.29 (1.90)	18.45 (3.19)	**60.33 (2.34)**	**61.01 (2.52)**	77.86 (5.56)	36.40 (13.90)
		ReSelfAttn	25.45 (3.40)	23.08 (3.20)	36.65 (4.65)	12.53 (2.34)	11.90 (2.25)	19.62 (3.58)	58.36 (3.02)	59.62 (2.85)	74.33 (5.53)	21.72 (17.19)
		SeqConv	27.06 (3.03)	24.66 (2.90)	38.78 (4.14)	14.15 (2.60)	13.51 (2.49)	22.16 (4.21)	58.37 (2.42)	58.91 (2.11)	73.13 (6.32)	13.69 (12.27)
		MOTIM	**22.71 (2.64)**	**20.38 (2.71)**	**33.71 (3.78)**	**10.25 (1.77)**	**9.75 (1.62)**	**17.55 (3.38)**	59.88 (2.38)	60.71 (2.44)	**78.52 (4.08)**	**43.65 (14.56)**
	1.0	Linear	23.34 (2.58)	20.95 (2.41)	34.72 (3.96)	10.86 (2.04)	10.45 (1.87)	18.42 (3.64)	58.83 (2.24)	59.90 (2.33)	76.89 (4.61)	41.23 (12.04)
		MoGRU	22.58 (2.51)	**20.06 (2.26)**	33.79 (3.77)	10.55 (2.10)	10.01 (1.93)	17.87 (3.56)	60.59 (3.08)	61.79 (3.25)	**78.66 (4.73)**	35.88 (18.16)
		ReSelfAttn	24.88 (2.87)	22.61 (2.80)	35.77 (3.90)	11.94 (2.45)	11.36 (2.37)	18.56 (3.64)	**60.64 (2.92)**	**61.98 (3.16)**	74.51 (5.83)	19.99 (15.53)
		SeqConv	25.17 (3.32)	22.90 (3.17)	36.39 (4.61)	12.26 (2.50)	11.77 (2.43)	19.34 (3.80)	60.44 (2.81)	60.98 (2.91)	74.29 (5.79)	19.39 (14.39)
		MOTIM	**22.46 (2.51)**	20.26 (2.55)	**33.22 (3.66)**	**10.24 (1.71)**	**9.85 (1.57)**	**17.27 (3.16)**	59.94 (2.48)	61.00 (2.70)	78.07 (3.93)	**41.94 (13.41)**
UMT1DM	-	*Repeat*	28.13 (6.07)	25.93 (5.64)	40.64 (9.19)	12.79 (4.03)	12.58 (4.01)	20.30 (6.45)	-	-	72.73 (9.30)	50.26 (25.43)
	0.0	Linear	27.08 (5.95)	25.28 (5.71)	37.77 (8.31)	13.61 (4.55)	13.22 (4.41)	21.20 (7.60)	53.64 (3.25)	56.14 (3.08)	73.87 (10.57)	45.43 (22.75)
		MoGRU	25.80 (6.20)	23.98 (6.20)	36.65 (8.56)	12.43 (3.92)	12.08 (3.81)	**19.87 (6.19)**	55.09 (3.97)	**57.16 (3.93)**	**74.71 (9.62)**	48.55 (22.53)
		ReSelfAttn	28.21 (5.90)	26.37 (5.61)	39.23 (8.40)	13.92 (3.80)	13.45 (3.73)	20.84 (5.68)	53.92 (3.93)	55.90 (4.06)	69.79 (12.62)	42.12 (28.48)
		SeqConv	29.33 (6.81)	27.51 (6.28)	40.12 (9.67)	16.33 (5.72)	15.73 (5.37)	24.21 (9.01)	53.88 (3.90)	55.89 (2.92)	70.17 (8.75)	27.95 (22.04)
		MOTIM	**25.40 (6.18)**	**23.47 (6.14)**	**36.16 (9.05)**	**12.41 (4.59)**	**12.03 (4.52)**	19.96 (7.70)	**55.24 (4.65)**	57.09 (4.73)	73.04 (19.17)	**52.51 (24.32)**
	1.0	Linear	25.74 (5.62)	23.94 (5.55)	36.60 (8.13)	13.02 (4.61)	12.70 (4.57)	21.10 (7.55)	54.57 (4.18)	56.74 (3.54)	**74.59 (10.02)**	46.99 (19.01)
		MoGRU	25.34 (5.71)	23.48 (5.66)	36.18 (8.16)	12.58 (4.71)	12.10 (4.75)	20.54 (7.62)	55.75 (4.67)	57.39 (4.31)	73.97 (10.62)	**52.75 (18.62)**
		ReSelfAttn	27.30 (5.85)	25.41 (5.54)	38.13 (8.62)	13.77 (4.68)	13.22 (4.54)	20.74 (7.23)	**56.14 (3.24)**	**57.58 (3.21)**	73.30 (9.74)	39.68 (23.75)
		SeqConv	28.31 (6.14)	26.37 (5.80)	39.55 (8.73)	15.59 (5.07)	15.11 (5.08)	23.62 (7.43)	54.57 (4.10)	56.42 (3.67)	73.08 (10.03)	29.49 (16.92)
		MOTIM	**25.21 (5.60)**	**23.47 (5.62)**	**35.61 (7.95)**	**12.57 (4.83)**	**12.03 (4.65)**	**20.48 (8.17)**	54.94 (4.52)	57.07 (4.32)	73.10 (16.33)	43.25 (25.33)

and false alarm rate, as they tend to achieve extreme values and present high variance. Instead, in our glucose excursion prediction task, we measure the performance of all individual values in the excursion being abnormal glycemia. More specifically, we filter all individual values in the excursion into binary labels using the hyperglycemia ($\geq$ 180 mg/dL) and hypoglycemia ($\leq$ 70 mg/dL) thresholds and evaluate the Matthews correlation coefficient (MCC) which takes value in $[-1, 1]$ following previous studies [29], [30]. In this evaluation, we exclude any patient from calculating the reported scores if and only if the following conditions hold for the patient: 1) the ground truth is all negative, and 2) the model correctly predicts them. Under such circumstances, no MCC can be calculated as there does not exist two classes.

4) Efficiency: Apart from the predictive performance, we also evaluate the efficiency of the methodologies, which is considered important when running the model on smart phones or low-power devices. In particular, we measure the following metrics: number of multiply–accumulate operations (MAC) in inference, number of model parameters, model inference time, and GPU memory consumption.

D. Benchmarks

To enable comparison to the state of the art, we include a naive *Repeat* baseline, a Linear baseline and several existing benchmarks from literature. Note that except for the naive *Repeat*, all methods are in the category of historical-value models. The *Repeat* baseline simply repeats the last known glucose value as the predicted excursion, thus no learning is involved at all. We include the Linear baseline to enable comparison with the MOTIM whose backbone is also based on linear connections. More specifically, Linear learns a linear mapping $f(\mathbf{g}_{t-I+1:t}) = \mathbf{W}\mathbf{g}_{t-I+1:t} + \mathbf{b}$ to directly map the past glucose values to the predicted excursion, where $\mathbf{W} \in \mathbb{R}^{H\times I}$ and $\mathbf{b} \in \mathbb{R}^{H}$ are its trainable parameters. For benchmarks from literature, we include multi-output GRU (MoGRU) [12], sequence-to-sequence convolution (SeqConv) method [15], and recurrent self-attention model (ReSelfAttn) [4] into the comparison to enable coverage on the approaches of encoding, namely recurrent, convolutional, and self-attention based encoding.

E. Implementation Details

The codes are implemented based on PyTorch, and the experiments are conducted on an NVIDIA RTX3090 GPU. For the historical-value benchmarks, we set the input length to 60 minutes and empirically adopt the model specifications in the original studies. For the MOTIM, we conduct a hyperparameter search for the lookback length in [1, 3, 5, 7] times the length of the prediction horizon. In training the models, we apply mean squared error (MSE) as the loss function $\mathcal{L}(\cdot, \cdot)$, and we constantly leave out the last 20% of the training data as the validation set. Detailed training setups are summarised in Table V.

F. Results

According to our personalised training experiment results in Table II, we confirm on the effectiveness of this MOTIM method as it achieves comparable performance to existing historical-value models. For example, in numeric performance of RMSE and MAPE and the abnormal glycemia performance, the MOTIM achieves the best or the second best in comparison with the benchmarks. Although the MOTIM is not at the first rank in TP scores, all the methods exhibit a degree of proximity to one another, with a marginal difference of $\sim 2\%$.

The above trend of the MOTIM excelling in its effectiveness also hold in our cross-instance adaptation experiments

TABLE IV
EFFICIENCY EVALUATION RESULTS. THE INFERENCE TIME IS AVERAGED OVER 1000 RUNS ON DIFFERENT MINI-BATCHES.

	MAC	Parameter	Time	GPU Memory
MoGRU	1.824G	2.373M	15.88(0.74)ms	1409MiB
ReSelfAttn	5.438G	593.665K	65.72(1.57)ms	1417MiB
SeqConv	4.008G	183.745K	37.98(1.24)ms	1861MiB
MOTIM	**786.432K**	**32.896K**	**2.96(1.76)ms**	**1207MiB**

shown in Table III. Apart from that, it is noticeable that the MOTIM converges faster by vertically comparing results in $PoU = \{0.0, 1.0\}$. The performance gap of the 2 settings of the MOTIM is smaller than other methods (e.g., less than 1mg/dL overall RMSE decrease (22.71 to 22.46) in the MOTIM *v.s.* around 2mg/dL for SeqConv (27.06 to 25.17) in OhioT1DM). This suggests the effectiveness of the MOTIM easing the learning difficulty by changing the sequence-level distribution learning to value-level, since the MOTIM is shown to rely less on the amount of personalised training data.

In addition to analysing the two tables individually, we make more observations with a cross-examination between Table II and III as follows: First, we compare Table II with the $PoU = 0.0$ experiments in Table III. This corresponds to a comparison that the amount of training data is fixed to 1 patient's amount, while the source of training data is different (personalised data *vs.* data from another patient). We observe a consistently better performance in the personalised experiment when comparing the two experiments entry by entry. Moreover, we observe that the performance gap in this comparison is different among the listed methods. More specifically, the MOTIM achieves smaller performance gaps than existing historical-value models (e.g., the performance gap of the first metric RMSE (all) is 22.71-22.44=0.27 mg/dL while the same entry of SeqConv is 27.06-25.95=1.11 mg/dL). This observation empirically demonstrates that first, the cross-instance variance indeed exists and causes learning difficulty as the personalised training consistently shows better performance, and second, the MOTIM is more robust to cross-instance change as the performance gap is narrower than the existing historical-value models.

Our second observation is drawn from comparing Table II and the $PoU = 1.0$ experiments of the MOTIM in Table III. In this comparison, both experiments end with a same training round using full personalised data, while the cross-instance adaptation is preceded with a pre-training phase using data from a different patient. As a result, we observe similar performance (e.g., the MOTIM performance gap of RMSE (all) between the two experiments is only 22.46-22.44=0.02 mg/dL). This indicates the feasibility of putting the MOTIM in a pre-training & fine-tuning pipeline, the pre-training phase using another patient's data does not pollute the following personalised training phase and the two experiments end up with similar convergence.

As shown in Table IV, the MOTIM is demonstrated to be significantly more efficient than other historical-value alternatives in terms of runtime and memory usage. Training the MOTIM consumes at least 200MiB GPU memory less than all other methods. In terms of the model structure, the MOTIM contains less than 20% parameters than the lightest counterpart and only 0.01% MACs in the feedforward computation. Compared to MoGRU, which achieves relatively better predictive performance than other historical-value models, the MOTIM contains only 1.3% the number of the MoGRU's parameters and achieves more than 10 times faster inference. Thus, it is concluded that the MOTIM achieves significantly higher efficiency than existing methodologies.

TABLE V
EXPERIMENTAL SETUP DETAILS.

Setup of	Details
MOTIM	layer_size: 128 inr_layers: 1 n_fourier_feats: 256 λ: 0.0 σ^2 scales: [0.01, 0.1, 1, 5, 10, 20, 50, 100]
MoGRU	layer_size: 512 num_layer: 2 bidirectional: False
ReSelfAttn	layer_size: 128 num_layer: 3 n_heads: 4 d_ff: 512
SeqConv	layer_size: 64 num_layer: 3
OhioT1DM	anchor_patients: [540, 544, 552, 567, 584, 596, 559, 563, 570, 575, 588, 591]
UMT1DM	anchor_patients: [2, 5, 7, 9, 11, 15, 17, 19, 21, 27, 30, 33]
Training	epochs: 50 batch_size: 64 optimiser: Adam learning_rate: 1e-3 gradient_clipping: 10.0 early_stopping_patience: 7 epochs

V. DISCUSSION

A. Limitations

In this study, we address the problem as a uni-variate forecasting issue, in line with the approaches in [12], [31]–[33]. As a pioneering study using time-index modeling, we exclude additional modalities such as administered insulin and carbohydrate intake to create a more controlled and pure experimental environment, considering the challenges with accurate data collection in real world. Incorporating exogenous inputs might enhance performance, but is beyond the scope of this investigation.

B. Future Work

We hope that the novel modeling based on time-index discussed in this paper could bring the research beyond the boundary of traditional historical-value modeling. For example, one essential breakthrough brought by the high efficiency of the MOTIM is that the receptive field of past values has been largely increased (Figure 1). A promising future research topic is using the MOTIM to fit to a longer local pattern (e.g.,

to a half-day or daily level) which we are empirically more confident in its personal uniqueness assuming sufficient data are available.

VI. Conclusion

In this research, we target the problem of data-driven glucose excursion prediction. We propose for the first time to change the existing historical-value modeling paradigm to the meta-optimised time-index modeling for this problem. We evaluate the methodology MOTIM in both its predictive performance and efficiency in a variety of application scenarios. Our experiments on the OhioT1DM and UMT1DM datasets demonstrate that the MOTIM achieves predictive performance comparable to the state of the art, while the efficiency of this methodology is significantly more efficient. Therefore, we conclude that the MOTIM is a promising new direction in the field of glucose prediction.

References

[1] D. Rodbard, "Continuous glucose monitoring: a review of successes, challenges, and opportunities," *Diabetes T & Therapeutics*, vol. 18, no. S2, pp. S2–3, 2016.

[2] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, "A review of personalized blood glucose prediction strategies for T1DM patients," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 33, no. 6, p. e2833, 2017.

[3] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, "Learning deep time-index models for time series forecasting," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[4] R. Cui, C. Hettiarachchi, C. J. Nolan, E. Daskalaki, and H. Suominen, "Personalised short-term glucose prediction via recurrent self-attention network," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2021, pp. 154–159.

[5] M. De Bois, M. A. El Yacoubi, and M. Ammi, "Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people," *Computer Methods and Programs in Biomedicine*, vol. 199, p. 105874, 2021.

[6] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *NPJ Digital Medicine*, vol. 4, no. 1, p. 109, 2021.

[7] W. Seo, S.-W. Park, N. Kim, S.-M. Jin, and S.-M. Park, "A personalized blood glucose level prediction model with a fine-tuning strategy: A proof-of-concept study," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106424, 2021.

[8] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 193–204, 2022.

[9] S. Langarica, M. Rodriguez-Fernandez, F. Núñez, and F. J. Doyle III, "A meta-learning approach to personalized blood glucose prediction in type 1 diabetes," *Control Engineering Practice*, vol. 135, p. 105498, 2023.

[10] L. B. Godfrey and M. S. Gashler, "Neural decomposition of time-series data for effective generalization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 2973–2985, 2017.

[11] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *CEUR Workshop Proceedings*, vol. 2675. NIH Public Access, 2020, p. 71.

[12] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, and J. Wiens, "Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1387–1395.

[13] J. Xie and Q. Wang, "Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge." in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data@IJCAI-ECAI*, 2018, pp. 97–102.

[14] T. El Idrissi, A. Idri, I. Kadi, and Z. Bakkoury, "Strategies of multi-step-ahead forecasting for blood glucose level using LSTM neural networks: A comparative study." in *The 13th International Conference on Health Informatics*, 2020, pp. 337–344.

[15] S. M. A. Zaidi, V. Chandola, M. Ibrahim, B. Romanski, L. D. Mastrandrea, and T. Singh, "Multi-step ahead predictive model for blood glucose concentrations of type-1 diabetic patients," *Scientific Reports*, vol. 11, no. 1, p. 24332, 2021.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[17] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] S. B. Taieb, A. Sorjamaa, and G. Bontempi, "Multiple-output modeling for multi-step-ahead time series forecasting," *Neurocomputing*, vol. 73, no. 10-12, pp. 1950–1957, 2010.

[20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[21] M. M. H. Shuvo and S. K. Islam, "Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration," *IEEE Journal of Biomedical and Health Informatics*, 2023.

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

[23] J. Daniels, P. Herrero, and P. Georgiou, "A multitask learning approach to personalized blood glucose prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 436–445, 2021.

[24] R. Cui, C. J. Nolan, E. Daskalaki, and H. Suominen, "Jointly predicting postprandial hypoglycemia and hyperglycemia using continuous glucose monitoring data in type 1 diabetes," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2023.

[25] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7537–7547, 2020.

[26] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7462–7473, 2020.

[27] L. Bertinetto, J. F. Henriques, P. H. Torr, and A. Vedaldi, "Meta-learning with differentiable closed-form solvers," *arXiv preprint arXiv:1805.08136*, 2018.

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

[29] S. Oviedo, I. Contreras, C. Quiros, M. Giménez, I. Conget, and J. Vehi, "Risk-based postprandial hypoglycemia forecasting using supervised learning," *International Journal of Medical Informatics*, vol. 126, pp. 1–8, 2019.

[30] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi, "Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning," *Health Informatics Journal*, vol. 26, no. 1, pp. 703–718, 2020.

[31] K. Zarkogianni, K. Mitsis, E. Litsa, M.-T. Arredondo, G. Fico, A. Fioravanti, and K. S. Nikita, "Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring," *Medical & Biological Engineering & Computing*, vol. 53, pp. 1333–1343, 2015.

[32] G. Alfian, M. Syafrudin, J. Rhee, M. Anshari, M. Mustakim, and I. Fahrurrozi, "Blood glucose prediction model for type 1 diabetes based on extreme gradient boosting," in *IOP Conference Series: Materials Science and Engineering*, vol. 803. IOP Publishing, 2020, p. 012012.

[33] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, "Blood glucose prediction with variance estimation using recurrent neural networks," *Journal of Healthcare Informatics Research*, vol. 4, pp. 1–18, 2020.

Jointly Predicting Postprandial Hypoglycemia and Hyperglycemia Using Continuous Glucose Monitoring Data in Type 1 Diabetes

Ran Cui[1], Christopher J Nolan[2], Elena Daskalaki[1], and Hanna Suominen[1,2,3]

***Abstract*— The development of continuous glucose monitoring (CGM) systems has enabled people with type 1 diabetes mellitus (T1DM) to track their glucose trajectory in real-time and inspired research in personalised glucose prediction. In this paper, our aim is to predict postprandial abnormal-glycemia events. Different from prior research which focuses on hypoglycemia only, we make the first attempt to establish our problem as the joint prediction of hyperglycemia and hypoglycemia. On this basis, we propose a machine learning model that learns from the pattern of 1 hour past glucose and makes predictions for the two tasks simultaneously using a unified backbone. Key benefits of our methodology include 1) requiring only the CGM sequence as the input, thus making it more widely applicable than other counterparts using extra inputs such as the nutrition details, and 2) minimising the computational cost as the two tasks are unified into a single model. Our experiments on the openly available OhioT1DM dataset achieve state-of-the-art performance (Matthew's correlation coefficient of 0.61 for hyperglycemia and 0.48 for hypoglycemia). To encourage further study, we release our codes at `https://github.com/r-cui/PostprandialHyperHypoPrediction` under the MIT license.**

I. Introduction

Type 1 diabetes mellitus (T1DM) is a chronic metabolic disorder affecting millions of people globally. Many people with T1DM wear continuous glucose monitoring (CGM) devices to constantly track their glucose trajectory [1], and personalised glucose prediction using machine learning (ML) on CGM data has drawn significant research attention [2], [3].

To date, the most straightforward setting of glucose prediction is short-term glucose prediction [4]–[9]. In this setting, a model is trained and evaluated using the full set of CGM data, and the learning target is to accurately predict the glucose value at a fixed horizon such as 30 or 60 minutes using recent glucose-related information as the input. However, glucose presents different patterns during the day. As shown in Figure 1, under the effect of food, postprandial glucose usually has a pattern from rising to declining that differs from other times such as the nocturnal glucose, which is relatively stable. Table I shows a statistical comparison of postprandial 4-hour glucose with glucose at other times using the OhioT1DM dataset [10]. The two-sample Kolmogorov–Smirnov (KS) tests [11] on all patients achieve a p-value less than 0.05, indicating their difference in probability distribution. This evidence suggests the potential of glucose prediction in a finer granularity, and, in this study, we specifically target the postprandial scenario.

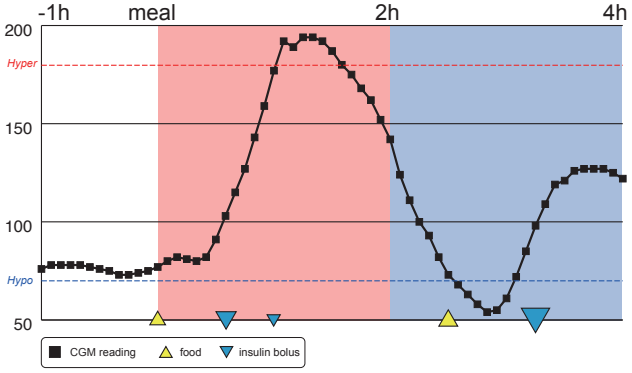


Fig. 1. A snippet of T1DM patient's real-life CGM data around a meal. This example shows a typical postprandial glucose pattern: the digestion of food causes glucose to peak, followed by its falling to a valley under the effect of natural metabolism and meal-corresponded insulin. Consequently, hyperglycemia and hypoglycemia are both risky in postprandial times.

Prior research has mainly approached postprandial glucose prediction from two directions, namely prediction at meal time [12]–[15] and prediction in fixed horizon [16], [17]. Meal-time prediction considers a pre-defined meal effective time range such as 4 hours after the meal, and all predictions regarding the time range are made at once at the meal time. This scheme reflects the application that the patient could re-plan the meal advised by the prediction result, e.g., adjust the planned food quantity or insulin amount. Although the envisaged application is of great value, a natural disadvantage of this scheme is that it cannot take factors after the meal intake into consideration. As illustrated in Figure 1, other meals and insulin intake events could be present at the patient's wish in real life, rendering the meal-time prediction inaccurate under such circumstances. In contrast, the fixed prediction horizon scheme considers a shorter prediction horizon such as 30 or 60 minutes, and the prediction is rolling starting right after the meal time until it covers all the meal effective time range. This scheme corresponds to the application that once a meal event is announced by the patient, the algorithm constantly runs to predict glucose for a short future, such that later events that reflect on glucose profile can be accommodated. Although having a shorter prediction horizon than the meal-time prediction scheme, the application of fixed horizon prediction is more flexible.

[1]School of Computing, The Australian National University, Canberra, Australia
[2]School of Medicine and Psychology, The Australian National University, Canberra, Australia
[3]Department of Computing, University of Turku, Turku, Finland
ran.cui, christopher.nolan, eleni.daskalaki, hanna.suominen@anu.edu.au

TABLE I

DISTRIBUTION IN THE FORMAT OF MEAN(STD) OF POSTPRANDIAL 4-HOUR CGM DATA AND OTHER CGM DATA IN THE OHIOT1DM DATASET, AND TWO-SAMPLE KOLMOGOROV–SMIRNOV (KS) TESTS PER PATIENT.

Patient ID	540	544	552	584	596	559	563	570	575	588	591
Postprandial 4h CGM	139.7 (60.7)	183.5 (59.5)	145.3 (55.7)	186.1 (64.2)	147.9 (51.1)	173.8 (72.8)	149.8 (51.3)	181.1 (66.2)	137.3 (62.3)	163.5 (55.3)	151.2 (57.9)
Other CGM	136.0 (53.2)	142.5 (52.7)	147.2 (54.0)	195.4 (65.7)	142.7 (46.0)	155.1 (66.6)	142.8 (48.3)	192.8 (56.7)	142.0 (56.9)	164.2 (47.7)	158.1 (58.2)
KS Test $p < 0.05$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In this study, we aim to design and develop a strategy for the prediction of postprandial abnormal-glycemia events under the fixed horizon prediction scheme. Specifically, we propose a problem formulation under the fixed prediction horizon scheme in which the prediction target is set to both postprandial hyperglycemia and hypoglycemia instead of only one of them as in previous studies [12], [13], [17]. The two targets are annotated simply by two distinctive configurations of time zones and thresholds (Figure 1) according to clinical suggestions [18], [19], so this expansion does not add extra complexity to the problem. Under our problem formulation, we propose a long short-term memory (LSTM) [20] based model that uses the 1-hour past CGM pattern to learn to predict the near future. Instead of separately obtaining two models for the two tasks (hypo-, hyper-glycemia), our method uses a single LSTM backbone to handle the temporal feature extraction for both tasks, thus the need for computational resources in both training and inference is minimised. As the envisaged user scenario is that the user notifies the system of entering the postprandial stage after each meal and the system constantly runs the prediction, we consider this optimisation important for running on mobile or low-power devices. As a reflection of the recent initiative of data sharing in this field [21], we evaluate our proposed methodology on the openly available OhioT1DM dataset [10] collected in everyday settings. Though there exist differences in the data used by different studies, our experiments achieve state-of-the-art results compared to existing literature. Moreover, to further study the effect of the aforementioned glucose distribution shift in postprandial scenario, we conduct an ablation experiment in which we train the model on the whole CGM dataset. The resulting performance deterioration and 10-fold increase of training process indicate that treating postprandial glucose prediction separately is indeed a valid approach that should be further pursued.

The contributions of our study are summarised as follows:

- To the best of our knowledge, we are the first to formulate the problem of jointly predicting postprandial hyperglycemia and hypoglycemia in the field of postprandial glucose prediction.
- We propose a ML approach that unifies the two prediction tasks into one single model so that the need for computational resources is minimised. It achieves state-of-the-art prediction quality compared to existing studies.
- Our statistical analysis and ablation study verify the impact of the glucose distribution shift problem in postprandial glucose prediction.

II. RELATED WORKS

Our study sits in the cross-domain of *short-term glucose prediction* and *postprandial glucose prediction*. Thus, we review related existing research in both fields.[1]

A. Short-Term Glucose Prediction

We refer to *short-term glucose prediction* [4]–[9] as the research field of predicting the glucose value at a fixed prediction horizon ranging from 15 to 120 minutes using recent glucose-related information as input, including CGM recordings, carbohydrates intake and insulin delivery. To date, most of these studies were evaluated on openly available datasets such as OhioT1DM [10] in purpose of finding better methodology enabled by open performance comparison. Deep learning models have been recently shown to be widely effective on this problem, such as recurrent neural networks (RNN) [9], dilated convolution neural networks (CNN) [7], multi-scale LSTM [8], and self-attention [6].

However, research on this topic was mostly designed and evaluated using all available CGM data. As discussed in the introduction, the unique distribution and pattern of postprandial glucose suggest the potential of distinguishing different scenarios in a day which is inadequately investigated in this field.

B. Postprandial Glucose Prediction

We refer to *postprandial glucose prediction* [12]–[17] as the research field of glucose prediction or hyperglycemia (hypoglycemia) prediction that targets only the postprandial scenario, using either only CGM recording or combine with meal information such as the nutrition components. Research in this field can be classified into two categories differing in whether the prediction is made at the meal time at once or is constantly made with a fixed prediction horizon. For the first category, [12], [13] studied meal-time prediction of postprandial hypoglycemia in 4 hours after the meal and their state-of-the-art performance was 0.48 in Matthew's correlation coefficient. Different in the prediction target, [14] studied meal time prediction of the full excursion of 2 hours postprandial glucose, and [15] studied the meal-time prediction of several features such as the area under curve of 2 hours postprandial and glucose value at 1 hour postprandial that could reveal the future trend of postprandial glucose. However, a natural inadequacy of meal-time prediction is that it is unable to accommodate a dynamic future, such as another meal or extra insulin intake in the meal effective time range as the patient may wish. On the other hand, the second category in

[1]For more comprehensive reviews of glucose prediction, we refer the readers to [2], [3].

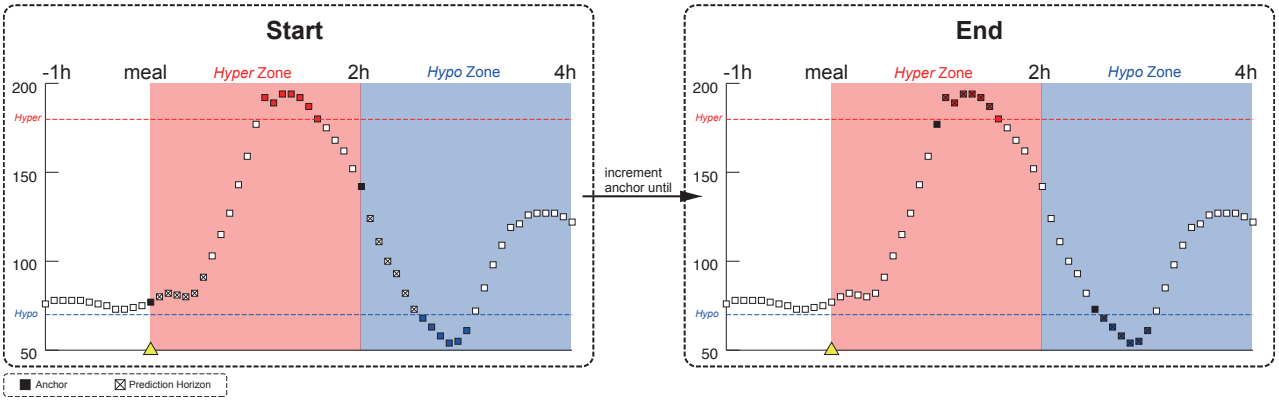


Fig. 2. The illustration of our examples extraction. Taking hyperglycemia for illustration, an anchor started to increment from the first timestamp in the red zone, and a corresponding example was generated at each incrementing step until it arrived just before the occurrence of a hyperglycemia event. The extraction of hypoglycemia was of a similar manner as illustrated in the blue zone.

this field follows the constant prediction strategy using fixed prediction horizon. In this category, [16] studied the use of nutrition absorption models in predicting the glucose value at 1 hour. [17] explored prediction of postprandial hypoglycemia in 30 minutes using random forest classification model and achieved 0.9 sensitivity and 0.9 specificity with a cost of 0.7 false alarm rate.

Our study is in relation to short-term glucose prediction as we follow the fixed prediction horizon strategy, but with a focus on postprandial data only. Moreover, our study is in line with the second category of postprandial glucose prediction, but targeting the joint prediction of postprandial hyperglycemia and hypoglycemia, and relying on only the CGM sequence which circumvents extra burdens such as estimating meal carbohydrates [22], [23].

III. Proposed Method

We present our formulation of postprandial hyperglycemia and hypoglycemia prediction in Section III-A, and the rest of Section III constitutes the description of our ML methodology for this problem.

A. Problem Formulation

Under the configuration of the sampling rate being 5 minutes, we set our prediction horizon to be 30 minutes (6 data points). In accordance with clinical consensus [18], [19], we consider 0-2 hours and 2-4 hours after the meal to be the effective time range for the two tasks of postprandial hyperglycemia and hypoglycemia, respectively. Under this configuration, the 4-hour postprandial glucose trajectory, denoted by $\mathbf{g} = [g_1, g_2, ..., g_{48}]$, was our source of examples extraction for a meal at the time of sample g_0. Specifically, for each task, we applied a moving anchor i starting at the first index of its time range. A new example (i, y_i) was generated each time the anchor incremented until the prediction horizon moved to the last 6 data points in the time range. To this

Algorithm 1 Pseudo Codes for Extracting Train/Test Examples from a Meal.

Input: $\mathbf{g}$, t ▷ CGM sequence, meal time
Initialise $ph = 6$ ▷ prediction horizon: 0.5 h

Case 1: hyperglycemia
Initialise $thres = 180$
Initialise $T = [1, 2, ..., 18]$ ▷ 0-2 h
procedure ExtractData($\mathbf{g}$, t)
  $data = []$
  for i in T **do**
    if $\mathbf{g}_{t+i} \geq thres$ **then**
      break
    end if
    if two consecutive in $\mathbf{g}_{(t+i+1):(t+i+ph)} \geq thres$ **then**
      $y = True$
    else
      $y = False$
    end if
    push tuple (t, i, y) to $data$
  end for
  return $data$
end procedure

Case 2: hypoglycemia
Initialise $thres = 70$
Initialise $T = [25, 26, ..., 42]$ ▷ 2-4 h
procedure ExtractData($\mathbf{g}$, t)
  Change $\geq$ to $\leq$ in line 10 and 13.
end procedure

end, the anchor i took a default range in

$$\text{hyperglycemia} \quad \{i \in \mathbb{N} : 1 \leq i \leq 18\}, \tag{1}$$

$$\text{hypoglycemia} \quad \{i \in \mathbb{N} : 25 \leq i \leq 42\}. \tag{2}$$

We defined that a hyperglycemia (hypoglycemia) event occurred if at least two consecutive data points in the prediction horizon crossed the threshold of 180 mg/dL (70 mg/dL) [24]. Consequently, the label y_i was a binary value determining by if such event occurred in the prediction horizon $\mathbf{g}_{i+1:i+6}$.

On this basis, an early stopping criterion in the incremental

process of the anchor i was applied: the increment of i early stopped at the last point before an event of interest being present (Figure 2). This was to reflect that the prediction corresponding to a certain meal was no longer required once the meal already caused a hyperglycemia (hypoglycemia). By repeating the procedure described above on all meals, the full postprandial dataset could be constructed using the raw personalised CGM data. Such dataset was in line with the fixed-horizon prediction research as reviewed in Section II-B. A pseudo code snippet for this training and testing examples generation is shown in Algorithm 1.

B. Classification to Regression

Under this problem formulation, a central idea of our methodology was to change the natural binary classification setup adopted by previous studies [12], [13], [17] into an approximated regression scheme. More specifically, instead of directly making a prediction of the binary y_i, we let our model learn to predict a numerical value of the max (min) glucose value in the prediction horizon, then the binary decision was made via the threshold 180 mg/dL (70 mg/dL). Although this translation was not exactly equivalent to the definition of two-consecutive points crossing the threshold, we adopt this scheme for its stronger supervision to the model training, because the binary supervision would lose the information of the severity of the hyperglycemia (hypoglycemia) event. Moreover, as the events tend to be rare or even do not present at all (especially for hypoglycemia), the example distribution tends to be significantly imbalanced. Our regression scheme also naturally avoided the need of dealing with the imbalanced class problem, such as choosing a proper class weighting in classification loss.

C. Unified Model and Joint Training for Hyperglycemia and Hypoglycemia

As the definition of hyperglycemia and hypoglycemia only differed in the threshold and time range, we approached the two tasks simultaneously using a unified model structure which took the recent glucose pattern as input. To be specific, we used the 1-hour recent CGM sequence $\mathbf{g}_{i-11:i}$ as the input and left out all other factors such as the meal and insulin amount in consideration of the difficulty of obtaining reliable meal amount estimation [25] and that the data may not be sufficiently large to capture the personalised effect of insulin. To encode the sequential feature, we fed the input 1-hour CGM sequence into an LSTM [20] backbone with hidden dimension D followed by a Rectified Linear Unit (ReLU) activation function [26]. The last hidden state $\mathbf{h}$ of the LSTM, given by

$$\mathbf{h} = \max(\text{LSTM}(\mathbf{g}_{i-11:i}), 0) \in \mathbb{R}^D, \tag{3}$$

was regarded as the encoded feature vector of the input CGM sequence. The LSTM backbone was followed by two linear heads, which took the identical $\mathbf{h}$ as the input, but were respectively responsible for the prediction of hyperglycemia and hypoglycemia. To reflect our regression scheme described in Section III-B, the two linear heads projected $\mathbf{h}$ to two numerical outputs that respectively corresponded to the maximum and minimum glucose in the prediction horizon. The two predictions were given by

$$\hat{g}_{\text{hyper}} = \mathbf{W}_{\text{hyper}} \cdot \mathbf{h} + b_{\text{hyper}} \in \mathbb{R}, \tag{4}$$

$$\hat{g}_{\text{hypo}} = \mathbf{W}_{\text{hypo}} \cdot \mathbf{h} + b_{\text{hypo}} \in \mathbb{R}, \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^D$ and $b \in \mathbb{R}$ represented the canonical linear weight and bias term in a linear projection, respectively.

Finally, mean squared error (MSE) loss between the predicted value and the ground truth was employed in supervising the model training, given by

$$\mathscr{L}_{\text{hyper}} = \frac{1}{M}\sum^{M}(\hat{g}_{\text{hyper}} - \max \mathbf{g}_{i+1:i+6})^2, \tag{6}$$

$$\mathscr{L}_{\text{hypo}} = \frac{1}{N}\sum^{N}(\hat{g}_{\text{hypo}} - \min \mathbf{g}_{i+1:i+6})^2, \tag{7}$$

where M and N denote the total number of training examples extracted for hyperglycemia and hypoglycemia, respectively.

As the two linear heads shared the same hidden state $\mathbf{h}$ as the input, the LSTM was encouraged to learn the temporal trend of the recent glucose without bias towards either task, such as predicting a high glucose value for hyperglycemia or a low glucose value for hypoglycemia. To promote this in the training process, in each training iteration, a hyperglycemia batch and a hypoglycemia batch were independently fed to the model, and the model was alternatively updated using the corresponding loss (i.e., $\mathscr{L}_{\text{hyper}}$ or $\mathscr{L}_{\text{hypo}}$, depending on the task of that batch). As will be shown in our ablation study in Section IV-E, our jointly designed methodology for the two tasks hyperglycemia and hypoglycemia achieved comparable performance to the default option of independently training two models for the two tasks, yet half of the computational resources was saved.

IV. Experiments

A. Dataset

We evaluated our proposed method for predicting postprandial hyperglycemia and hypoglycemia using the publicly available dataset OhioT1DM [10] created by the US Ohio University. This dataset contained eight weeks of data from CGM devices, self-reported insulin intake information and meal events for 12 T1DM patients. We made use of the CGM recordings and followed the train/test split officially defined by the author of the dataset in evaluating our method, in which the test set comprised around ten days of data for each patient[2]. Ethical approval (2020/411) was obtained from the Human Research Ethics Committee of The Australian National University to use de-identified data for this study.

B. Evaluation Method

In evaluating the performance of our proposed methodology, we evaluated the root mean square error (RMSE) for our regression output in accordance with existing studies in general glucose prediction, and further evaluated the sensitivity (SE),

[2]In the raw dataset, the patient with ID 567 provided no meal information in the test split, so we excluded this patient from our experiments.

TABLE II

MAIN PERFORMANCE COMPARISON IN THE FORMAT OF MEAN (STD) OVER 10 REPEATED EXPERIMENTS WITH DIFFERENT RANDOM SEEDS. BOLD SCORES WERE STATISTICALLY SIGNIFICANTLY BETTER ($p < 0.05$) THAN ALL BASELINES VIA THE UNPAIRED T TEST.

	Hyperglycemia						Hypoglycemia					
	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑
Short-Term - [4]	19.08	-	-	-	-	-	19.08	-	-	-	-	-
Short-Term - [5]	18.93	-	-	-	-	-	18.93	-	-	-	-	-
Short-Term - [6]	17.97	-	-	-	-	-	17.97	-	-	-	-	-
Postprandial - [17]	-	-	-	-	-	-	-	0.90	0.91	0.70	-	25.5
								(0.03)	(0.02)	(0.04)		(1.97)
Postprandial - [13]	-	-	-	-	-	-	-	0.69	0.8	-	0.24	-
Postprandial - [12]	-	-	-	-	-	-	-	0.71	0.79	-	0.48	-
[17] Replicated	-	0.83	0.81	0.58	0.50	19.28	-	0.82	0.94	0.85	0.34	19.4
		(0.01)	(0.00)	(0.01)	(0.00)	(0.23)		(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
Baseline - Dummy	27.01	0.03	1.00	0.00	0.15	5.00	17.44	0.05	1.00	0.40	0.17	5.00
Baseline - AdaBoost	22.03	0.71	0.88	0.46	0.53	16.53	17.42	0.07	1.00	0.70	0.13	10.25
	(0.07)	(0.01)	(0.01)	(0.01)	(0.01)	(0.28)	(0.14)	(0.02)	(0.00)	(0.09)	(0.03)	(2.36)
Baseline - Random Forest	19.45	0.58	0.94	0.33	0.55	14.65	14.14	0.31	1.00	0.48	0.39	8.93
	(0.04)	(0.01)	(0.00)	(0.01)	(0.01)	(0.19)	(0.02)	(0.01)	(0.00)	(0.01)	(0.01)	(0.23)
Baseline - MLP	18.79	0.57	0.94	0.33	0.55	14.79	13.23	0.38	0.99	0.51	0.42	8.94
	(0.21)	(0.01)	(0.00)	(0.01)	(0.01)	(0.28)	(0.19)	(0.05)	(0.00)	(0.04)	(0.02)	(1.54)
Ours	**18.23**	0.66	0.95	0.33	**0.61**	15.01	13.25	0.48	0.99	0.50	**0.48**	10.97
	(0.35)	(0.04)	(0.01)	(0.03)	(0.02)	(0.98)	(0.17)	(0.05)	(0.01)	(0.05)	(0.02)	(0.89)

specificity (SP), false alarm rate (FA) which were calculated out of the confusion matrix for the final classification. We also included Matthew's correlation coefficient (MCC) as an overall reflection of the classification result. Additionally, to evaluate the timeliness of our prediction, we calculated the detection time (DT) of the start of a hyperglycemia or hypoglycemia event averaged over the meals that were correctly predicted by our algorithm. In these metrics, MCC took values in $[-1,1]$ to reflect the overall classification quality and others in $[0,1]$, larger (smaller) values reflected better performance for SE, SP, and MCC (FA). For DT, the theoretically optimal value was 30 minutes as it was our preset prediction horizon.

C. Implementation Details

Due to the small size of the patient cohort, we trained personalised model for each patient in our experiments. The full training and evaluation were implemented using PyTorch. Normalisation was applied to all CGM data using mean and standard deviation estimated via training data. In the model configuration, we applied a uni-layer LSTM with hidden dimension $D = 50$, resulting in the full model containing around 10k trainable parameters. The model was tuned using Adam optimiser [27] with batch size 64 and learning rate of 0.001 under a gradient norm clipping at 1.0. To illustrate the computational efficiency under this configuration, each training step took around 0.01 seconds, and the full training pipeline for all 11 patients in OhioT1DM dataset took around 4 minutes on an Apple M1 CPU. As the amount of personalised postprandial CGM data was limited, instead of using hold-out training data as a validation set and conduct model selection, we trained our model on the full training data and fixed the training epochs to 50. To enable statistical significance testing, we repeated all experiments 10 times with different random seeds, and reported the averaged scores and their standard deviation in all tables.

D. Results

As shown in Table II, we compared our proposed method to state-of-the-art studies in the *short-term glucose prediction* field (Section II-A) and the *postprandial glucose prediction* field (Section II-B). Moreover, the main body of our comparison was our own implementation of several baselines using traditional ML algorithms.

a) Comparing to Short-Term Glucose Prediction Studies: As shown in the first section of Table II, the studies [4]–[6] were all conducted using the same OhioT1DM dataset while evaluated using RMSE between the actual and predicted glucose value without including the other metrics used in our study as the prediction of abnormal-glycemia events was out of their scope. For RMSE, our study was slightly different to these studies as our regression target was the max and min glucose in the next 30 minutes instead of the exact glucose at 30 minutes ahead. However, our methodology achieved a comparable RMSE to these studies when predicting hyperglycemia and an RMSE significantly better than these studies when predicting hypoglycemia, suggesting a performance no lower than the methodologies proposed in these studies if applied to our task.

b) Comparing to Postprandial Glucose Prediction Studies: Among the postprandial glucose prediction studies in our performance comparison in the second section of Table II, [17] followed the fixed prediction horizon scheme with a 30-minute prediction horizon, while [12], [13] adopted the meal-time prediction scheme. These studies were conducted on datasets unavailable to the community due to restrictions of the clinical trials, which made our comparison less straightforward. However, though a lower DT was presented, our study achieved the state-of-the-art MCC, indicating an promising overall performance of our prediction.

c) Comparing to Our Implemented Baselines: To enhance the comparative assessment, we replicated the methodology proposed in [17] and implemented several classical

TABLE III

ABLATION EXPERIMENTS IN THE FORMAT OF MEAN(STD) OVER 10 REPEATED EXPERIMENTS WITH DIFFERENT RANDOM SEEDS.

	Hyperglycemia						Hypoglycemia					
	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑	RMSE↓	SE↑	SP↑	FA↓	MCC↑	DT↑
Individual Training	18.64	0.64	0.96	0.29	0.62	14.22	13.50	0.40	1.00	0.47	0.45	10.13
	(0.14)	(0.01)	(0.01)	(0.02)	(0.01)	(0.28)	(0.08)	(0.05)	(0.00)	(0.05)	(0.03)	(0.79)
Binary Supervision	-	0.80	0.80	0.60	0.47	18.11	-	0.70	0.90	0.90	0.24	15.7
		(0.02)	(0.01)	(0.01)	(0.01)	(0.54)		(0.06)	(0.05)	(0.04)	(0.06)	(0.80)
Short-Term Training	19.67	0.51	0.97	0.25	0.57	11.71	12.88	0.30	1.00	0.52	0.37	7.13
	(0.21)	(0.02)	(0.01)	(0.02)	(0.01)	(0.38)	(0.15)	(0.04)	(0.01)	(0.09)	(0.03)	(0.98)
Ours	18.23	0.66	0.95	0.33	0.61	15.01	13.25	0.48	0.99	0.50	0.48	10.97
	(0.35)	(0.04)	(0.01)	(0.03)	(0.02)	(0.98)	(0.17)	(0.05)	(0.01)	(0.05)	(0.02)	(0.89)

ML algorithms as shown in the third section of Table II. All results here were based on the same dataset and experimental conditions. To be specific, we first implemented a dummy model, which simply treated the most recent known glucose value as the regression output $\hat{g}_{\text{hyper}}$ and $\hat{g}_{\text{hypo}}$. As the dummy model had no ability to learn anything, no 10-fold experiment was conducted. This model was used to set up the theoretical lower bound of performance and to show that other methods had the learning ability to an extent. On this basis, we implemented several traditional regression algorithms, namely random forest (RF), AdaBoost, and multi-layer perceptron (MLP) using the same 1-hour past CGM sequence as input, and further replicated the methodology proposed by [17] which was an RF model using explicit features calculated out of the past CGM. In comparison with these baseline strategies, our methodology achieved the highest MCC and lowest RMSE, and comparable performance on other metrics.

d) Discussion: In the OhioT1DM dataset, postprandial hyperglycemia happened around 5 times more often than hypoglycemia which coincides with hypoglycemia usually causing more severe and acute symptoms than hyperglycemia, leading to T1DM patients being more cautious in preventing hypoglycemia. The more significant class imbalance problem in hypoglycemia data caused an overall worse performance in hypoglycemia prediction task (0.13 in MCC). However, this performance gap implied a possibility of having hypoglycemia prediction performance comparable to hyperglycemia prediction by collecting more data as the two tasks only differed in the threshold and time range. Besides, we observed the positive correlation between SE and FA and the trade-off between FA and DT. These observations were understandable because a higher SE indicated a higher tendency of the model to predict the presence of an event, which would also increase the FA. Since the DT was determined by the earliest detection of the event, the more often predicted events would also achieve a better DT. We achieved an average DT of 15 minutes for hyperglycemia and 11 minutes for hypoglycemia. Due to that rapid-acting insulin can be effective in 10-15 minutes [28], and the 20-minute expected recovery time of hypoglycemia [29], we consider our achieved DT to be a promising performance in practical application.

E. Ablation Studies

To further evaluate the validity of our method, we conducted three ablation studies targeting on different arguments in our method (Table III).

a) Individual Training: This ablation experiment aimed at investigating the validity of our joint model for hyperglycemia and hypoglycemia against treating the two tasks separately. Thus, instead of training the model alternatively using data from the two tasks, we trained and evaluated two models individually. In comparing our main experiment to the first row of Table III, our unified model achieved results close to the individual training scheme in all metrics. This observation suggested the capability of the shared LSTM backbone to properly encode the temporal feature of the input glucose history which was then used for prediction on the two tasks. By unifying the two tasks into one model, we minimised the need for computational resources both in inference time and memory requirement.

b) Binary Supervision: Different from previous studies [12], [13], [17], our method approached the classification output in the way of translating the predicted max or min glucose in the prediction horizon to binary result using the defined thresholds. To investigate its effectiveness, we conducted the ablation study of applying binary supervision. More specifically, we changed the output module of our model to a binary classification layer, and the training was supervised by cross-entropy loss with class weights estimated via the class distribution in training data. As shown in the second row of Table III, though the DT was increased, this training strategy led to almost double FA and caused a significant drop of overall predictive performance ($\sim$12% MCC), indicating the numerical supervision being more effective.

c) Short-Term Training: In this study, we treated postprandial glucose prediction in separate to general glucose prediction due to the distribution shift from postprandial glucose to overall glucose as shown in Figure 1 and Table I. To further investigate the effectiveness of this decision, we conducted the ablation experiment of training the model using examples extracted from full CGM data which was consistent to the setting of short-term glucose prediction, yet the testing data were kept unchanged as only postprandial. As can be seen from the third row of Table III, this change not only did not help on increasing the overall performance, but rather caused a deterioration in both the MCC and DT. Additionally, the full training data were approximately of 10 times larger amount than postprandial training data, meaning that invoking a large amount of less relevant training data had only misled the learning.

V. Limitation and Future Work

One limitation of this study is that we did not include meal-related factors such as carbohydrate amount and insulin dosing into our model. Discussing the trade-off between 1) simpler input leads to better applicability, and 2) more complex input provides more potential for better performance, is a good topic for future research. Moreover, future research in direction of a long-term learning model that continuously improves itself using newly collected data from daily life is of great value.

VI. Conclusion

This study identifies the distribution shift of glucose data in the postprandial scenario, and focuses on the problem of postprandial hyperglycemia and hypoglycemia prediction. For the first time we formulate the problem of joint prediction of postprandial hyperglycemia and hypoglycemia. On this basis, we propose a unified ML model that handles the two tasks together and can learn from the glucose pattern only without the need for additional inputs such as meal nutrition in detail. Our experiments on the OhioT1DM dataset achieve state-of-the-art capability to predict postprandial hyperglycemia and hypoglycemia in comparison with existing studies.

Acknowledgements

This research has been delivered in partnership with and funded by Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and healthcare providers. We gratefully acknowledge the funding from the ANU School of Computing for the first author's PhD studies.

References

[1] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, S. Garg, L. Heinemann, I. Hirsch, S. A. Amiel *et al.*, "International consensus on use of continuous glucose monitoring," *Diabetes Care*, vol. 40, no. 12, pp. 1631–1640, 2017.

[2] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, "A review of personalized blood glucose prediction strategies for t1dm patients," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 33, no. 6, p. e2833, 2017.

[3] V. Felizardo, N. M. Garcia, N. Pombo, and I. Megdiche, "Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review," *Artificial Intelligence in Medicine*, vol. 118, p. 102120, 2021.

[4] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[5] T. Yang, X. Yu, N. Ma, R. Wu, and H. Li, "An autonomous channel deep learning framework for blood glucose prediction," *Applied Soft Computing*, vol. 120, p. 108636, 2022.

[6] R. Cui, C. Hettiarachchi, C. J. Nolan, E. Daskalaki, and H. Suominen, "Personalised short-term glucose prediction via recurrent self-attention network," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2021, pp. 154–159.

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

[8] T. Yang, R. Wu, R. Tao, S. Wen, N. Ma, Y. Zhao, X. Yu, and H. Li, "Multi-scale long short-term memory network with multi-lag structure for blood glucose prediction." in *5th International Workshop on Knowledge Discovery in Healthcare Data*, 2020, pp. 136–140.

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

[10] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *CEUR Workshop Proceedings*, vol. 2675. NIH Public Access, 2020, p. 71.

[11] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[12] S. Oviedo, I. Contreras, C. Quirós, M. Giménez, I. Conget, and J. Vehi, "Risk-based postprandial hypoglycemia forecasting using supervised learning," *International Journal of Medical Informatics*, vol. 126, pp. 1–8, 2019.

[13] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi, "Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning," *Health Informatics Journal*, vol. 26, no. 1, pp. 703–718, 2020.

[14] D. Adelberger, F. Reiterer, P. Schrangl, C. Ringemann, T. Huschto, and L. del Re, "Prediction of postprandial glucose excursions in type 1 diabetes using control-oriented process models," *IFAC-PapersOnLine*, vol. 54, no. 15, pp. 466–471, 2021.

[15] E. A. Pustozerov, A. S. Tkachuk, E. A. Vasukova, A. D. Anopova, M. A. Kokina, I. V. Gorelova, T. M. Pervunina, E. N. Grineva, and P. V. Popova, "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus," *IEEE Access*, vol. 8, pp. 219 308–219 321, 2020.

[16] R. A. Karim, I. Vassányi, and I. Kósa, "After-meal blood glucose level prediction using an absorption model for neural network training," *Computers in Biology and Medicine*, vol. 125, p. 103956, 2020.

[17] W. Seo, Y.-B. Lee, S. Lee, S.-M. Jin, and S.-M. Park, "A machine-learning approach to predict postprandial hypoglycemia," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–13, 2019.

[18] S. Daenen, A. Sola-Gazagnes, J. M'Bemba, C. Dorange-Breillard, F. Defer, F. Elgrably, E. Larger, and G. Slama, "Peak-time determination of post-meal glucose excursions in insulin-treated diabetic patients," *Diabetes & Metabolism*, vol. 36, no. 2, pp. 165–169, 2010.

[19] Y. Altuntaş, "Postprandial reactive hypoglycemia," *Şişli Etfal Hastanesi tıp Bülteni*, vol. 53, no. 3, p. 215, 2019.

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

[21] S. Cichosz, "Data sharing of continuous glucose monitoring data: The need for a new paradigm?" *Journal of Diabetes Science and Technology*, vol. 15, no. 4, p. 961, 2021.

[22] C. Hettiarachchi, E. Daskalaki, J. Desborough, C. J. Nolan, D. O'Neal, H. Suominen *et al.*, "Integrating multiple inputs into an artificial pancreas system: Narrative literature review," *JMIR Diabetes*, vol. 7, no. 1, p. e28861, 2022.

[23] C. Hettiarachchi, N. Malagutti, C. Nolan, E. Daskalaki, and H. Suominen, "A reinforcement learning based system for blood glucose control without carbohydrate estimation in type 1 diabetes: In silico validation," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 950–956.

[24] R. M. Bergenstal, A. J. Ahmann, T. Bailey, R. W. Beck, J. Bissen, B. Buckingham, L. Deeb, R. H. Dolin, S. K. Garg, R. Goland *et al.*, "Recommendations for standardizing glucose reporting and analysis to optimize clinical decision making in diabetes: the ambulatory glucose profile," *Diabetes Technology Society*, 2013.

[25] N. Brew-Sam, M. Chhabra, A. Parkinson, K. Hannan, E. Brown, L. Pedley, K. Brown, K. Wright, E. Pedley, C. J. Nolan *et al.*, "Experiences of young people and their caregivers of using technology to manage type 1 diabetes mellitus: Systematic literature review and narrative synthesis," *JMIR Diabetes*, vol. 6, no. 1, p. e20973, 2021.

[26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

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

[28] S. K. Garg, S. L. Ellis, and H. Ulrich, "Insulin glulisine: a new rapid-acting insulin analogue for the treatment of diabetes," *Expert Opinion on Pharmacotherapy*, vol. 6, no. 4, pp. 643–651, 2005.

[29] A. Nakhleh and N. Shehadeh, "Hypoglycemia in diabetes: An update on pathophysiology, treatment, and prevention," *World Journal of Diabetes*, vol. 12, no. 12, p. 2036, 2021.