Imperial College of Science, Technology and Medicine
Department of Electrical and Electronic Engineering
Centre for Bio-Inspired Technology

Deep Learning Methods for Improving Diabetes Management Tools

John Spencer Mbia Daniels

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London
October 2022

Declaration of Originality

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been appropriately referenced. Any contribution made to this research by others is explicitly acknowledged in the text.

Copyright Declaration

Abstract

Diabetes is a chronic disease that is characterised by a lack of regulation of blood glucose concentration in the body, and thus elevated blood glucose levels. Consequently, affected individuals can experience extreme variations in their blood glucose levels with exogenous insulin treatment. This has associated debilitating short-term and long-term complications that affect quality of life and can result in death in the worst instance. The development of technologies such as glucose meters and, more recently, continuous glucose monitors have offered the opportunity to develop systems towards improving clinical outcomes for individuals with diabetes through better glucose control. Data-driven methods can enable the development of the next generation of diabetes management tools focused on i) informativeness ii) safety and iii) easing the burden of management. This thesis aims to propose deep learning methods for improving the functionality of the variety of diabetes technology tools available for self-management.

In the pursuit of the aforementioned goals, a number of deep learning methods are developed and geared towards improving the functionality of the existing diabetes technology tools, generally classified as i) self-monitoring of blood glucose ii) decision support systems and iii) artificial pancreas. These frameworks are primarily based on the prediction of glucose concentration levels.

The first deep learning framework we propose is geared towards improving the artificial pancreas and decision support systems that rely on continuous glucose monitors. We first propose a convolutional recurrent neural network (CRNN) in order to forecast the glucose concentration levels over both short-term and long-term horizons. The predictive accuracy of this model outperforms those of traditional data-driven approaches. The feasibility of this proposed approach for ambulatory use is then demonstrated with the implementation of a decision support system on a smartphone application. We further extend CRNNs to the multitask setting to explore the effectiveness of leveraging population data for developing personalised models with limited individual data. We show that this enables earlier deployment of applications without significantly compromising performance and safety.

The next challenge focuses on easing the burden of management by proposing a deep learning framework for automatic meal detection and estimation. The deep learning framework presented employs multitask learning and quantile regression to safely detect and estimate the size of unannounced meals with high precision. We also demonstrate that this facilitates automated insulin delivery for the artificial pancreas system, improving glycaemic control without significantly increasing the risk or incidence of hypoglycaemia.

Finally, the focus shifts to improving self-monitoring of blood glucose (SMBG) with glucose meters. We propose an uncertainty-aware deep learning model based on a joint Gaussian Process and deep learning framework to provide end users with more dynamic and continuous information similar to continuous glucose sensors. Consequently, we show significant improvement in hyperglycaemia detection compared to the standard SMBG.

We hope that through these methods, we can achieve a more equitable improvement in usability and clinical outcomes for individuals with diabetes.

Acknowledgements

In my journey of completing this PhD I have come to appreciate the idea of the PhD not as a designation, but as a transformative journey. This not only encompasses the immensely rewarding trials and tribulations of failed attempts and eventual successes, but also the daily academic and extracurricular interactions littered throughout this period.

I would first and foremost like to thank my supervisor Prof. Pantelis Georgiou for constantly supporting me throughout this process. He granted me with free rein to pursue the research questions that I was interested in and shape the projects I embarked on, which made this experience an enriching pursuit. He also made himself available for advice, general feedback, and encouragement through the more disorienting periods in this journey whenever I sought guidance and for this I am truly grateful. I would also like to thank Dr. Pau Herrero and Dr. Kezhi Li for being great fountains of knowledge to turn to and for providing constructive critique on the ideas that form the foundations of this thesis.

Beyond this I would also like to my fellow colleagues in the Centre for Bio-Inspired Technology (CBIT). In particular, I would like to thank Mohamed El-Sharkawy, Peter Pesl, Chengyuan Liu, Taiyu Zhu, and Amparo Guemes for the stimulating conversations and activities in the diabetes group. The diligence and dedication to your research helped keep me inspired in my own work. Further gratitude is extended to Prof. Nick Oliver, Dr. Monika Reddy, Dr. Chukuwma Uduku, and the rest of the clinical team at St. Mary's Hospital for providing needed guidance where possible. I would also like to extend special gratitude to Izabela Wojcicka-Grzesiak and Wiesia Hsissen for expertly attending to the countless administrative queries I had during this time.

Finally, I would like to thank my family and friends for the endless love and encouragement to persevere. This has been crucial in reaching this milestone, and as such, I view this work as a triumph of that collective effort.

‘Doubt is the key to knowledge.’

Persian Proverb

Abbreviations

T1DM Type 1 Diabetes Mellitus

T2DM Type 2 Diabetes Mellitus

RMSE Root Mean Square Error

MARD Mean Absolute Relative Difference

MCC Matthews Correlation Coefficient

CGM Continuous Glucose Monitoring

GLUT Glucose Transporter

GP Gaussian Process

MGP Multitask Gaussian Process

SMBG Self Monitoring of Blood Glucose

EDA Electrodermal activity

NN Neural Network

ANN Artificial Neural Network

DNN Deep Neural Network

CNN Convolutional Neural Network

RNN Recurrent Neural Network

CRNN Convolutional Recurrent Neural Network

LSTM Long Short Term Memory

SVM Support Vector Machine

SVR Support Vector Regression

LVX Exogenous Latent Variable Model

ARX Exogenous Autoregressive Model

BGRI Blood Glucose Risk Index

HbA1c Haemoglobin A1c

PPGR Post Prandial Glycaemic Response

AP Artificial Pancreas

CSII Continuous Subcutaneous Insulin Infusion

MDI Multiple Daily Injections

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Diabetes is a chronic metabolic disease where the regulation mechanism of blood glucose concentration is compromised. As a result, affected people experience large variations in their blood glucose profile. This is typically characterised by persistent high blood glucose concentration (hyperglycaemia). The areas of primary concern include extended periods of hyperglycaemia, as well as extended periods of low blood glucose concentration (hypoglycaemia). These periods are generally referred to as adverse glycaemic events and are associated with worsening quality of life.

As one of four major non-communicable diseases, addressing challenges diabetes is a pressing endeavour in modern public health. Globally, it is estimated that over 400 million people are living with diabetes today [5]. This represents an increase from 1980 where the incidence of diabetes was 108 million to today, and is projected to increase further to 700 million in 2045 [6]. In addition, this increase is geographically distributed in unequal degrees, with a larger percentage increase in low- and middle-income countries (LMICs) which compounds the complexity of solutions to address this issue [6, 5].

The increasing trend in incidence of diabetes imposes a burden on global healthcare. Consequently, this calls for the introduction and democratisation of technology that is geared towards enabling long-term self-management of diabetes.

1.1 Motivation

Diabetes technology can be defined collectively as the constellation of hardware and software that empower people with diabetes to self-manage their condition, from blood glucose levels to informing lifestyle decisions [7]. Some of these devices include glucose meters, continuous glucose monitors (CGM), insulin pens, insulin pumps, and combined systems such as the artificial pancreas.

Studies in the literature on the application of diabetes technology have indicated that self management of diabetes is critical in maintaining a high standard in the quality of life of people with diabetes. The primary method of diabetes management is glucose control [8]. This is mainly carried out through self-monitoring of blood glucose (SMBG), where users prick their finger with a lancet and insert a blood sample in a glucose meter to obtain a blood glucose reading [9]. This gives a snapshot of the users' glucose profile and guides decision-making at particular points for ingesting snacks to avoid hypoglycaemia or take insulin to avoid hyperglycaemia [10]. Although SMBG method allows some management of glucose excursions, this is sparse and irregular, limited to the few moments that the measurements can be made, and can cause discomfort due to constant finger pricking. With this in mind, the emergence of wearable technology has offered an alternative approach through continuous and ambulatory monitoring which facilitates novel tools for tight glycaemic control [8, 11, 10, 12].

Wearable devices, including continuous glucose monitors (CGM) and beyond, offer the potential in improving diabetes management by facilitating continuous and minimally invasive monitoring of physiological signals. The continuous glucose monitor represents the state-of-the art in measuring glucose concentration levels and provides regular measures at five minute intervals. In daily living, situations such as physical activity and stress that are known to have an effect of the metabolic system [13, 14, 15].

The Cambrian explosion in wearable devices and the availability of smartphones has resulted in an associated increase in the data. With the emergence of machine learning, particularly deep learning, this data can be leveraged to develop personalised mobile decision support systems and artificial pancreas systems. The success of deep learning in multiple fields (eg. computer vision, image recognition, natural language processing, financial engineering, and wireless communication) has led to active research in various areas of healthcare.

In the field of diabetes technology, few studies have been explored that utilise deep learning and its complementary approaches to improve the functionality of diabetes management tools. This improvement in diabetes technology with deep learning can enable greater functionality and move diabetes management tools towards easing the burden of managing diabetes.

1.2 Research Objectives

The self management of diabetes is critical in maintaining a high standard in the quality of life of T1DM subjects. Wearable devices, such as CGM, offer the potential in improving diabetes management by facilitating continuous and minimally invasive monitoring of physiological signals.

The research detailed supports the thesis that deep learning approaches can leverage the data produced from these devices in order to extend the functionality of current diabetes management tools, and across the heterogeneous range of tools available. This serves to provide more information to the individuals living with diabetes and further ease the burden of managing diabetes.

This thesis mostly focuses on the diabetes population that is predominantly dependent on insulin therapy in order to maintain a blood glucose concentration level within a healthy range (euglycaemia).

In order to achieve this outcome, the project aims to fulfil the following objectives:

1. **To improve forecasting accuracy of glucose concentration levels with continuous glucose monitors.**

   In the current landscape of the artificial pancreas system and decision support systems, the recommendations/control actions typically depend on current glucose values. However, given the slow nature of subcutaneous insulin delivery/ingested carbs this can only minimise the periods spent outside the glycaemic target range. This objective looks at the using deep learning models to develop personalised models for users in order to pre-empt and potentially avoid adverse glycaemic events.

2. **To leverage population data in scarce data settings to develop glucose forecasting algorithms.**

   One of the outstanding challenges in the area of developing data-driven models for diabetes technology is the availability of data for training personalised models. Collection of data is generally costly and as a result can hamper the development and early deployment of data-driven systems for diabetes technology tools. Furthermore, generalised models trained on population data are not as effective given the existence of inter-individual variability given the complexity of the metabolic system. We aim to develop methods that effectively leverage population data to develop personalised models in the face of scarce data.

3. **To develop resource efficient models to ease the burden of announcing meals for improved glucose control.**

   Meals and snacks are a significant contributor to potential scenarios of postprandial hyperglycaemia in T1DM users. However, consistently having to log meals through carbohydrate counting means that for at random moments the participants miss logs for mealtimes or consequent snacks that AP systems are

unable to handle the resulting PPGR. We aim to ease the burden of the user by developing multitask learning methods of detecting and estimating unannounced meals using signals from continuous glucose monitors and insulin. This would further develop fully automated closed-loop insulin delivery, and reduce the burden on T1DM users.

4. **To develop algorithms to improve the limited information provided through self-monitoring of blood glucose.**

   Given the lack of homogeneity in the landscape of diabetes technology solutions, it is necessary to develop algorithms to cater to the different range of tools available. Currently, machine learning approaches leverage continuous glucose monitors that provide glucose level measurements at fixed intervals. However, a large portion of people with diabetes also rely on glucose meters that provide sparse and irregularly sampled and thus incompatible with systems that rely on continuous glucose methods. We aim to extend the current algorithms for continuous estimation and forecasting to people with diabetes using glucose meters.

1.3 Thesis Structure

This thesis is broadly composed of three sections that tackle improving three areas of diabetes management as seen in Fig 1. below.

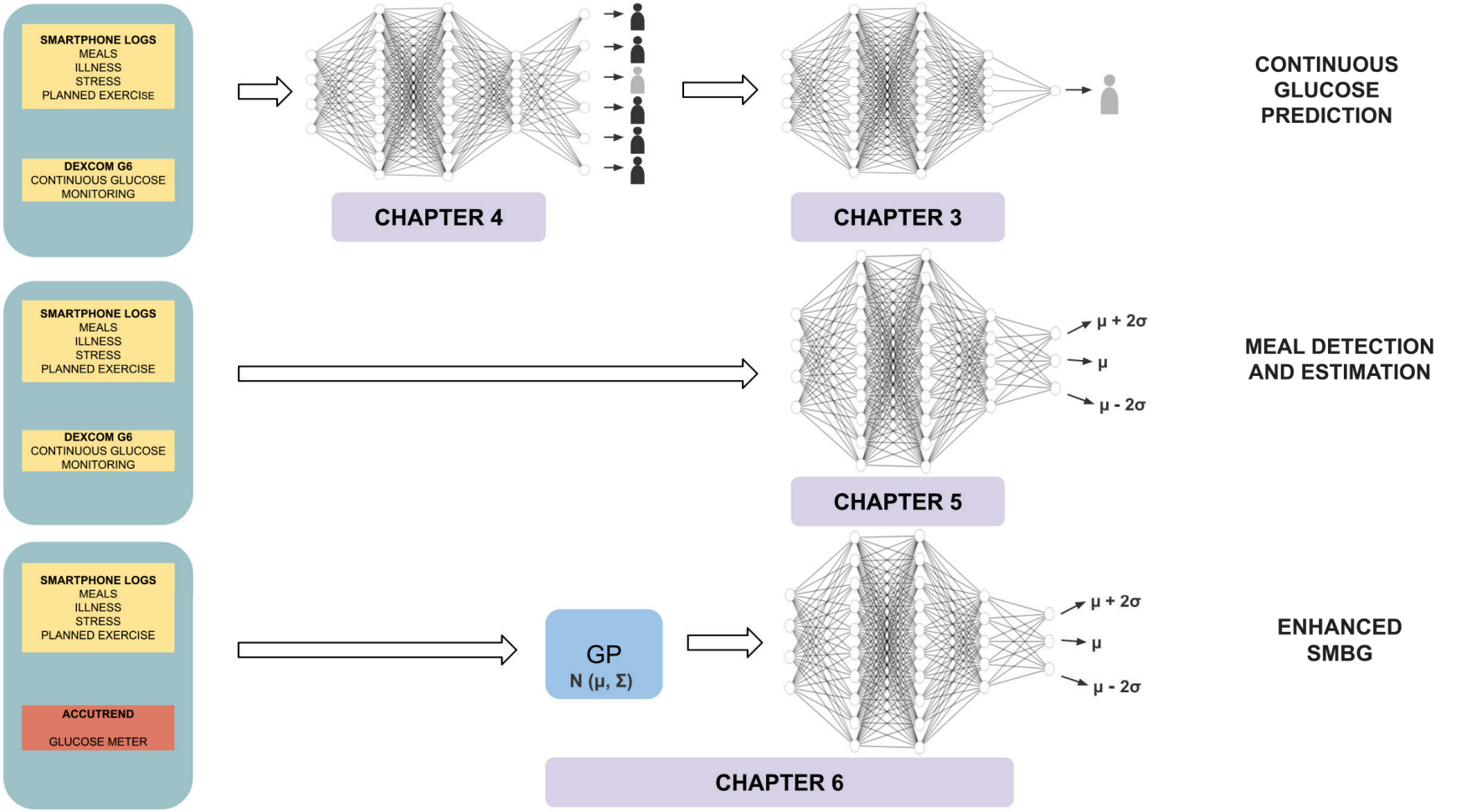

Figure 1.1: A simplified illustration of the thesis outline showing the general inputs considered from smartphones, CGMs, and insulin pumps, as well as the approaches for tackling formulated tasks. **Top:** Chapters 3 and 4 detail deep learning methods for the task of blood glucose prediction, and leveraging population data with multitask learning to develop personalised models. **Middle:** Chapter 5 details formulating multitask deep learning models for automatically detecting and estimating meals to ease diabetes management. **Bottom:** Chapter 6 involves work to enable monitoring of glucose concentrations levels from sparse glucose meter readings through uncertainty-aware deep learning methods.

Chapter 2: An Overview of Diabetes and Diabetes Technology

Chapter 2 gives the reader an overview of the current landscape of diabetes technology, focusing on the wearable devices and/or algorithms that enable diabetes self-management. This begins with background information on diabetes as a chronic metabolic disease and challenges and complications people with diabetes face in daily living. A review is then carried out on the sensors and devices that comprise state-of-the art in diabetes technology, the physiological signals and relevant variables considered, and the examples of algorithms powering the current diabetes management tools in use today.

Chapter 3: Deep Learning for Glucose Prediction

Chapter 3 addresses the challenge of forecasting glucose levels from continuous glucose monitors. We introduce the convolutional recurrent neural network (CRNN), a deep learning model, that automatically learns the features from multimodal signals to provide both short- and long-term forecasts. This model is evaluated on both synthetic and clinical datasets, and we show an improvement in performance over traditional baseline methods. The work from this chapter is previously published in J5 and C2 (*See List of Publication in Appendix A*).

Chapter 4: Multitask Learning for Personalised Glucose Prediction

Chapter 4 tackles the challenge of developing personalised deep learning models in the presence of scarce available data. We introduce a multitask learning approach to leverage data from other people with diabetes while maintaining personalisation. We present results that show this approach yields better performance over comparable learning strategies in scarce data settings and at both short- and long- term prediction horizons. We also demonstrate consistent model performance in hypoglycaemia detection with different training data sizes. This chapter is based on work

previously published in J4 and W1 (*See List of Publication in Appendix A*).

Chapter 5: Multitask Learning for Automatic Meal Detection and Estimation

Chapter 5 is focused on easing the burden people with diabetes face in daily living, particularly in tracking their carbs. Current artificial pancreas systems require manual announcement of meals which hampers true closed-loop insulin delivery. We develop a meal detection and estimation algorithm based on a multitask deep learning model to detect and estimate meals. We also investigate the performance of this approach in improving controllers towards a fully closed-loop artificial pancreas. This chapter is based on work previously published in J3 and C1 (*See List of Publication in Appendix A*).

Chapter 6: Uncertainty-Aware Learning for Enhanced SMBG

Chapter 6 provides a solution to the challenge of improving self-monitoring of blood glucose (SMBG) for individuals with diabetes use glucose meters instead of continuous glucose monitors (CGM). This allows low-cost alternatives such as glucose meters to provide continuous estimation and forecasting of glucose levels. We compare our approach to baseline machine learning approaches used in forecasting glucose concentration levels. Finally, the proposed approach is compared to the standard SMBG to evaluate the improvement in adverse glycaemic event detection.

Chapter 7: Conclusion and Future Work

This chapter concludes the works presented in this thesis. The original contributions are highlighted in the context of potential applications in diabetes technology. We then discuss outstanding limitations with a view towards directions for future work to overcome them.

Chapter 2

An Overview of Diabetes and Diabetes Management Tools

2.1 Physiology of Glycaemic Feedback and Diabetes

In human physiology, it is essential that biochemical variables are consistently kept in a narrow range to ensure adequate health. Homeostasis is the term that defines "the self-regulating process by which biological systems tend to maintain stability while adjusting to conditions that are optimal for survival" [16]. This is accomplished mainly through negative feedback. This project focuses on ensuring the continual regulation of the plasma glucose concentration, one of the more important physiological variables in the body, without increasing burden on the user.

Glycaemic homeostasis requires negative feedback to control the level of plasma glucose concentration in the blood. Plasma glucose is supplied by dietary intake and is used as a source of energy by the various tissues in the body. The acceptable range of plasma glucose concentration for normal glycaemic conditions is 3.9 - 10 mmol/L. In normal functioning human physiology, this is adequately carried out

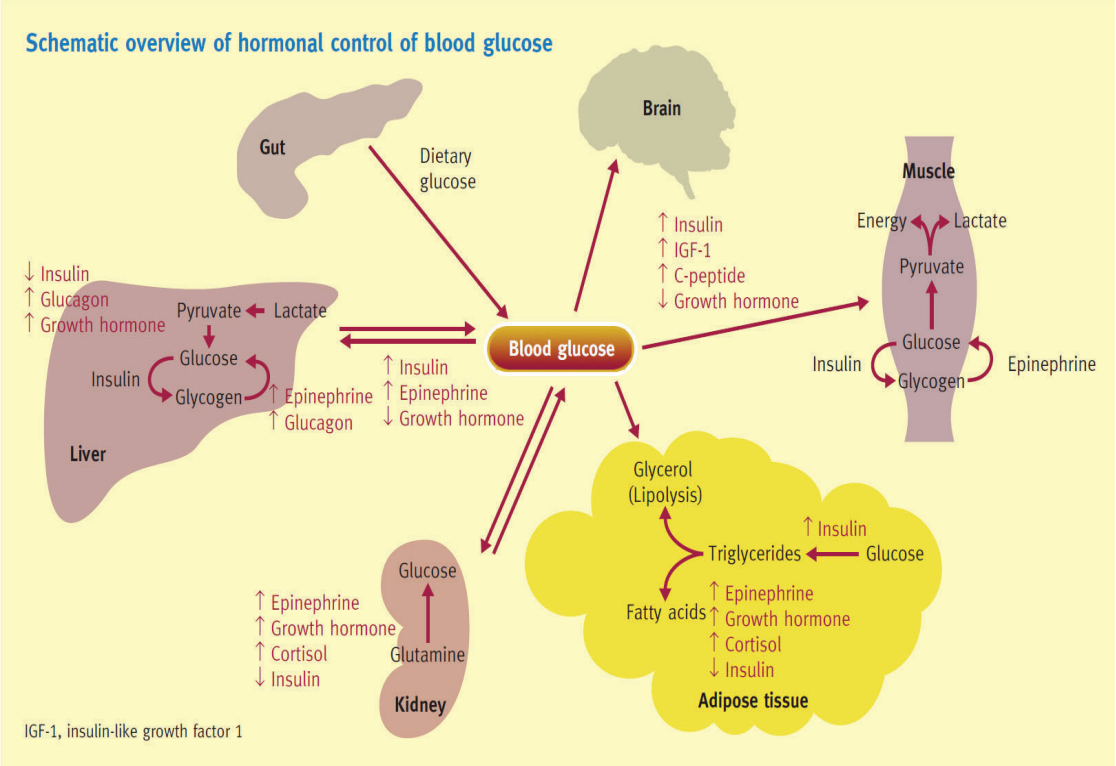

Figure 2.1: A figure showing the mechanism of hormonal control of blood glucose concentration. [1]

through monitoring by a number of biological mechanisms and hormonal regulation [1]. However, this natural regulation by the human body can be distorted leading to deviation from the acceptable range.

Hyperglycaemia is the condition characterised by the persistent elevated plasma glucose concentration above the upper range of 10.0 mmol/L (180 mg/dL) [17]. At the most extreme case, glucose concentration levels higher than 13.8 mmol/L (250 mg/dL), the subject with diabetes is at risk of diabetic ketoacidosis (DKA). This is a condition where there is a build-up of ketones that increase the acidity of blood [1]. In the long term, increased frequency of such adverse glycaemic events can lead to poor outcomes such as cardiovascular complications, loss of vision (diabetic retinopathy), kidney damage (nephropathy), nerve damage (neuropathy), and loss of limbs.

Conversely, hypoglycaemia is characterised by persistent plasma glucose concentration below the lower range of 3.9mmol/L (70 mg/dL) [18]. This can result in

functional impairments which is particularly dangerous during everyday activities such as driving. In extreme adverse scenarios where glucose concentration levels fall further to below 3mmol/L (54 mg/dL) this can result in a coma, and even death in worst cases.

In daily living, there are more activities outside meals and insulin that affect the glucose concentration level in myriad ways. This generally ranges from physical activity to undertaking stressful situations.

Physical activity is a recommendation for healthy living in the general population [19, 20]. Current guidelines for T1D care also incorporate recommendations for physical activity due to the benefits in managing diabetes. Exercise, from a physical standpoint, improves physical fitness, maintains a healthy weight, and improves insulin sensitivity. However, with these benefits come challenges in undertaking physical activity given the effect on glucose control [19].

The guideline recommends aerobic and anaerobic exercise for the benefits [13]. Aerobic exercise comprises low-moderate intensity activities such as running, rowing, and other cardio-related activities that typically targets a large set of muscles. The active muscles increase glucose uptake and thus along with increased insulin sensitivity can lead to hypoglycaemia. On the other hand, anaerobic exercise is considered to be high intensity activities such as resistance training, and high intensity interval training (HIITs). During the undertaking of such exercises, there is a change in neuroendocrine hormones that promotes an increase in the rate of conversion of glycogen from the liver to glucose into the bloodstream as shown in 2.1. Along with the typical recommendation of reduced insulin administration prior to exercise, this could lead to an increase in the glucose concentration level and may result in hyperglycaemia [13].

The glycaemic responses in daily living as a result of physical activity is more complex given the different scenarios that arise in undertaking exercise regimes as well as the metabolic state of the T1D individual. Firstly, as described earlier, glycaemic

response during and after exercise is mainly attributed to the type of exercise (i.e aerobic or anaerobic) as well as the intensity level and duration. In addition to this is a factor that Gassaleti *et al.* [20] attribute to 'metabolic memory' - the given course of glucose trajectory given past adverse glycaemic events. Prior hyperglycaemia correlates with increased inflammatory and oxidative stress responses, whereas, prior hypoglycaemia correlates with blunted autonomic responses and glucose counterregulation. As a result, extreme hyperglycaemia (>250mg/dL) at the onset of intense physical activity can potentially lead to diabetic ketoacidosis (DKA) and prior hypoglycaemia results in an increased likelihood of future hypoglycaemia events [19, 13].

Stress is also acknowledged as a potential contributor to hyperglycaemia in both healthy and diabetic populations [21, 22, 23, 15]. Kyrou *et al.* [24] show that exposure to stressful situations has a two part effect: the insulin sensitivity is depressed, and simultaneously, the glucose production is increased through the liver. Contrary to the non-diabetic population, this excess glucose production during a stress response is not metabolised in the aftermath of the stressor. This is supported by a number of studies that have shown that stress is strongly associated with poorer glycaemic control [25].

In addition to having an effect on glucose concentration levels, the hormonal changes due to these scenarios affect multiple organs as seen in Figure 2.1.

2.2 Sensors and Wearable Devices

In the last decade, the proliferation of sensors and low power systems has enabled continuous monitoring that has shown potential in medical applications. This adds to devices that have already existed in the diabetes technology ecosystem. In this section we review the various technologies that have been proposed, both as standalone and combined systems, as a result and are leveraged as tools for diabetes

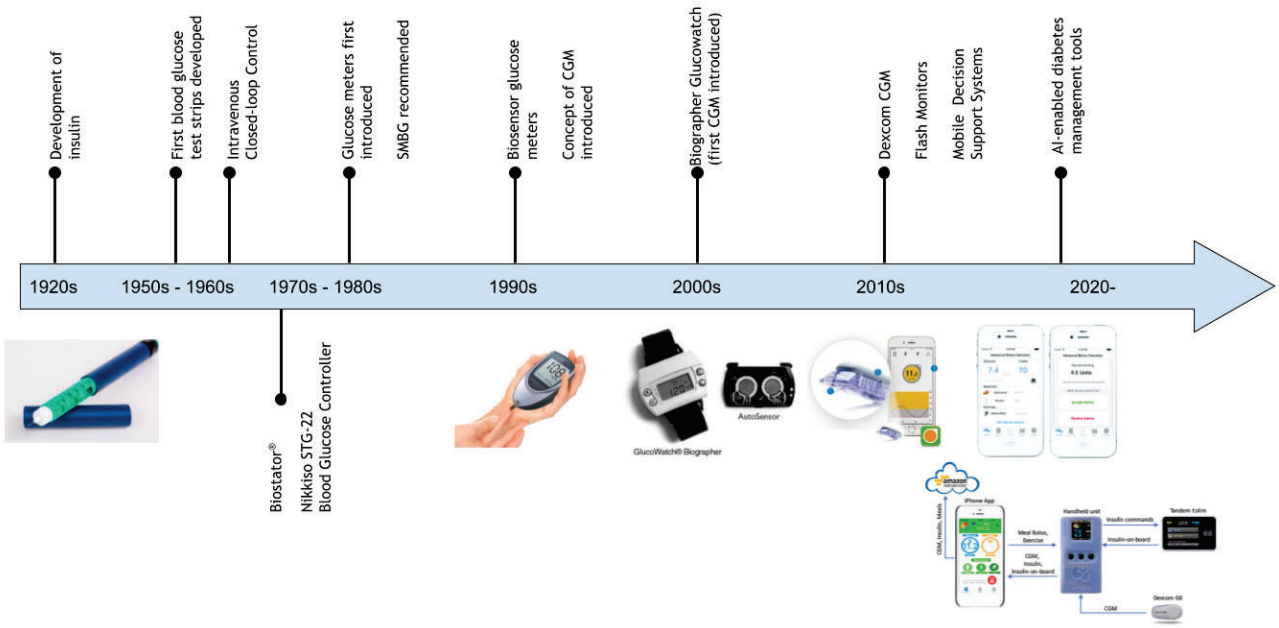

Figure 2.2: A timeline highlighting the pace of development of diabetes technology tools. This comprises sensors and devices for detecting glucose concentration level, insulin for controlling blood glucose levels, decision support systems and artificial pancreas systems. [2].

management.

2.2.1 Glucose Meters

The current glucose meter is the standard tool that is universally available to the Type 1 diabetes population for self-monitoring of blood glucose (SMBG). The glucose meter is a relatively simple and easy-to-use technology that combines electrochemistry and electronics [9]. The electrochemistry typically consists of test strips/biosensors containing enzymes that react with the user's whole blood sample - provided from the capillaries with a finger prick [26, 27]. The ensuing enzymatic reaction produces a tiny current that is converted to a calibrated glucose concentration reading with an A-to-D converter [26, 27].

The development of modern glucose meters can be traced to the Ames Reflectance Meter (1970) among other similar glucose meters, which mainly operated based on

photometry - where the intensity of the test strip colour was indicative of blood glucose concentration [9]. Over time the development of glucose meters towards the current form was ultimately dictated by the need for accuracy, portability, and ease of use. This led to the development of Accutrend, Glucotrend, Precision QiD, One Touch, and Glucocard II which are smaller handheld systems, easier to use, and more precise and accurate [9, 26]. Glucose meters enable assessing multiple measures of glucose levels throughout the day as needed for SMBG and MDI therapy.

Although multiple challenges have been resolved in moving towards consistent use throughout the day, there are still multiple factors that can interfere with the performance of the glucose meter. Primarily, the main source of error readings can be attributed to the user in relation to performing the testing [27]. This can be due to the manner in which the test strips are handled or the blood sample being below the sufficient amount required [27].

Environmental factors such as humidity can lead to rehydration of the enzymes which can lead to reduced reactivity [27]. Alternatively, extreme temperature differences can also lead to discrepancies in results owing to the subsequent effect on enzyme activity. Another consideration is the presence of drugs such as acetamenophin which can interfere with the accuracy of measurements. Lastly, the discrepancy in results from the true glucose concentration levels is much larger in the prandial and postprandial state (i.e. the period after meals are ingested) - a larger difference between capillary blood glucose concentration and venous blood glucose concentration - as compared to a fasting state [27].

These challenges with glucose meters inform the guidelines necessary for effective decision-making. When implementing the SMBG approach to glucose control it is advised to assess the blood glucose concentration levels i) prior to meals and snacks, ii) prior to exercise, iii) prior and during critical tasks such as driving , iv) at suspected hypoglycaemia events and after subsequent treatment, and finally v) at bedtime [7].

2.2.2 Continuous Glucose Monitors

CGM devices are low power, minimally invasive devices that make use of electrochemical sensors to provide blood glucose measures by inferring from the glucose concentration in the interstitial fluid [11, 3]. We provide insights into the operation of CGM, along with the challenges and opportunities towards improving diabetes management.

Figure 2.3: A Dexcom continuous glucose monitor showing the sensor(1), transmitter(2) and app interface(3) for viewing glucose time series [2].

The dominant type of CGM devices take the similar mode of operation as glucose meters regarding measuring glucose concentration via electrochemistry [12]. As seen in Figure 2.5, the sensor takes the form of a needle-like structure that is embedded just below the skin in subcutaneous tissue, with the abdomen or arm as the recommended application sites. The glucose concentration is sensed via a chemical reaction and monitoring the resulting electrical current. For some devices, calibration is required with 1-2 fingerstick measures everyday while more recent devices are self-calibrating and therefore can operate accurately without fingerstick measures [28, 3].

Continuous glucose monitoring (CGM) devices emerged as an improvement over SMBG as a tool for diabetes management [29, 30]. There is increasing evidence for

the clinical benefits in diabetes management given that readings are sampled every 5 minutes as opposed to actively at irregular checkpoints for SMBG [31]. As evident in Figure 2.4, using the glucose meter results in being unable to detect as many adverse glycaemic events as using the CGM. Furthermore, even with the detection of such adverse glycaemic events, the lack of general information about glucose dynamics or glucose trend from SMBG affects the outcomes of subsequent decision-making.

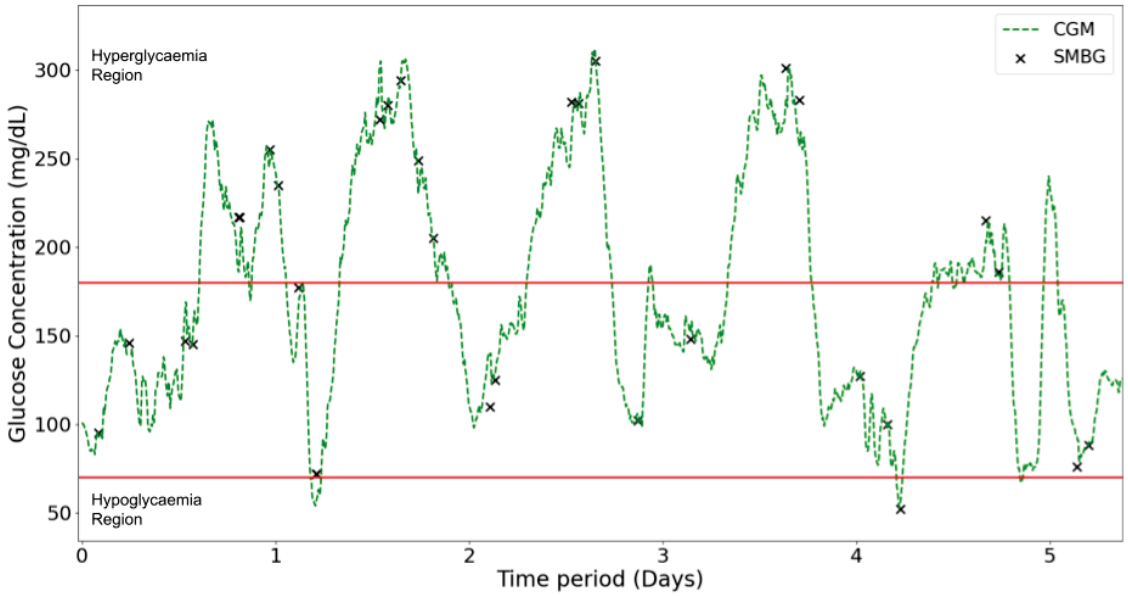

Figure 2.4: A comparison of the profiles of an Type 1 diabetes individual using a CGM and SMBG approaches. The CGM is sampled regularly and more frequently than SMBG. Consequently, adverse glycaemic events are mostly missed by adopting the SMBG monitoring instead of CGM monitoring. The number of adverse glycaemic events detected is also considerably less with SMBG monitoring relative to CGM monitoring.

The growing studies involved in assessing clinical value of continuous glucose monitoring have shown the clinical benefits of these devices. These studies have shown that increased use of CGM reduces the time spent out of advised range (70 - 180 mg/dL) [11, 10, 12, 31]. The inclusion of alarms for adverse glycaemic events helps to minimise their duration as corrective action can be taken. Alternatively, hypoglycaemia detection can be used in closed loop insulin delivery systems to suspend insulin delivery to enable quicker recovery [32].

The opportunities presented with the introduction of continuous glucose monitors also pose some challenges to their adoption. CGM only provides an estimate of the

blood glucose value from the glucose in the interstitial fluid. This leads to a time lag of 5-6 minutes due to the transport of glucose from blood vessels to the subcutaneous layer [11]. Furthermore, over time other measurement discrepancies exists due to drift and calibration errors. Similar to glucose meters, the accuracy of continuous glucose monitors (CGMs) are susceptible to interference; this can emanate from a variety of sources including compression resulting in compression artefacts, and drugs such as acetamenophin [12]. In addition, continued adoption is reported to hindered due to the alarm fatigue, potentially resulting from false alarms [10].

The use of machine learning models to enhance the functionality of the continuous glucose monitor (CGM) is a promising direction in the improvement of diabetes management [33]. However, further work would be required in order to fully realise the opportunities of CGMs in diabetes management systems.

2.3 Diabetes Management Tools

2.3.1 Artificial Pancreas

An artificial pancreas, also referred to as closed-loop insulin delivery, is one of the latest developments in diabetes technology for enabling tight glucose control in the Type 1 diabetes population [34, 35]. The development of insulin in 1920 kick-started this route to improving diabetes management for individuals living with diabetes [34]. However, this management is a delicate balance as there is a need to consider factors such as insulin-on-board (IOB), insulin-to-carb ratio (IC ratio), insulin sensitivity factor (ISF), and target blood glucose level. The artificial pancreas senses glucose levels, and infers the necessary insulin infusion using an algorithm that takes these factors into consideration to maintain a healthy glucose range.

The Biostater was developed in 1972 following intravenous closed-loop control in the 1960s [34]. This is the first recorded artificial pancreas and was primarily used

as a research tool and in inpatient settings. This device involved blood withdrawal for continuous glucose measurement and intravenous insulin infusion to achieve a set level of control. As a result, the lack of portable components restricted the use to the lab settings. The subsequent miniaturization and improvement of individual components over time as shown in Figure 2.2 has allowed the artificial pancreas to move out of the research setting in the lab and be commercialised in the real world.

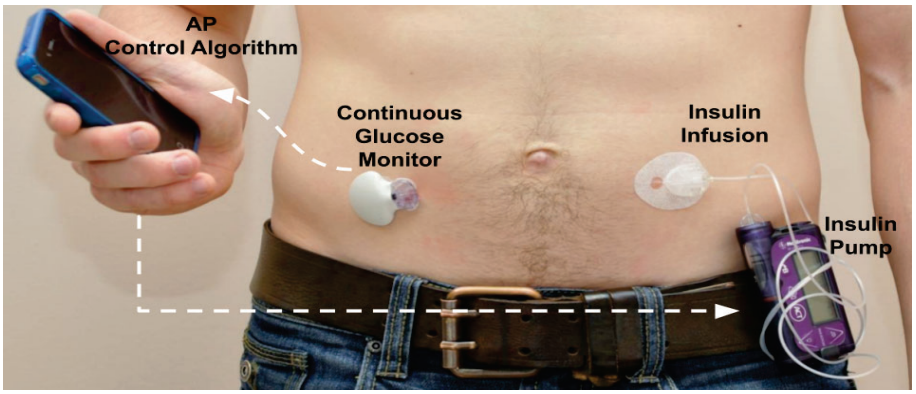

Figure 2.5: A typical current artificial pancreas system comprises a continuous glucose monitor which is the sensor(1), the smartphone (2) and app interface(3) for viewing glucose time series [2].

The current artificial pancreas consists of a CGM for frequently monitoring glucose concentration levels at regular intervals (5 minutes), an insulin pump for continuous subcutaneous insulin infusion, and a control algorithm running on a dedicated platform (i.e. a smartphone or custom hardware) or within the insulin pump for providing feedback and determining the insulin needed to maintain tight glycaemic control. The Medtronic Minimed 670G, Tandem Control IQ, and Diabeloop represent the current state-of-the-art in artificial pancreas technology for individuals with Type 1 diabetes that are commercially available [36]. However, these are not fully closed-loop system given that they deliver basal insulin automatically but require the user to input the carbohydrate content at meal times in order to estimate the appropriate insulin. This remains a challenge in attaining a fully closed-loop AP system [37, 36].

Nevertheless, the artificial pancreas has garnered attention and enthusiasm in the diabetes community. This can be seen in the development of open-source alternatives

(OpenAPS) developed within the community [38]. The technical, safety, and ethical considerations involved should spur researchers and all concerned stakeholders to work towards making a fully closed-loop artificial pancreas available.

2.3.2 Decision Support Systems

Decision support systems are software tools that have been developed in recent years to facilitate the ease in decision-making for people with diabetes. Earlier systems have focused on enabling clinicians in advising individuals on optimising their glucose control [39, 40, 41]. The advent of smartphones enabled the development of mobile applications that can allow individuals with T1D to have these decision support systems on-the-go.

The decision support systems have so far been used in the following categories:

1. Providing recommendations for adjustments to insulin delivery: In the first instance, we note that managing insulin delivery is a delicate procedure that takes a number of factors into account. The standard formula used for calculating the necessary insulin dose is provided below:

$$\begin{aligned} I &= I_{meal} + I_{correction}, \\ I &= \frac{Carbs(g)}{ICR} + \frac{BG(t) - BG_T}{ISF} - IOB, \end{aligned} \tag{2.1}$$

   where ICR is the insulin-to-carb ratio, ISF is the insulin sensitivity factor (also known as a correction factor), and IOB refers to the insulin on board. The insulin sensitivity factor is also a highly individualised parameter that quantifies the effect of 1 unit of rapid acting insulin in reducing blood glucose concentration levels over a period of 2-4 hours. BG(t) refers to the current blood glucose concentration level and BG_T refers to the target blood glucose level.

To this end, decision support systems have been developed based on this that are intended to reduce the accumulation of errors from misestimation and optimise the calculated insulin dose [42, 43, 44]. One scenario involves prandial insulin doses where an insulin bolus is calculated from an estimated carbohydrate size and an insulin-to-carb ratio (ICR). The insulin-to-carb ratio (ICR) is a highly individualised parameter that an individual with diabetes comes to after appropriate education and observing postprandial glucose patterns. This is shown in Equation 2.1. In addition, the insulin dose can also be given to correct a high blood glucose concentration level down to a target blood glucose concentration level, BG_T when in a fasting state.

2. Detecting unannounced meals and estimating carbohydrates: As noted previously, the insulin bolus for meals requires the individual with diabetes to estimate the carbohydrate size in a given meal and some decision support systems focus on optimising the insulin dose. However, this is an added burden and is missed occasionally which can lead to worsening outcomes [45, 46]. Decision support systems focusing on this have the objective of automatically detecting meals, and in some cases estimating carbohydrate size in order to enable postprandial glucose control [47, 48, 49, 50].

3. Prediction of adverse glycaemic events (i.e hypoglycaemia and hyperglycaemia): Alternatively, some decision support systems focus on predicting adverse glycaemic events in order to allow the individual with diabetes to have enough time to take action and avert the occurrence of this adverse event [51]. A number of systems in this case tend to focus on hypoglycaemia prediction, wherein the following situations are considered: short-term hypoglycaemia prediction[52], nocturnal hypoglycaemia prediction [53, 54], postprandial hypoglycaemia prediction [42, 55], and post-exercise hypoglycaemia [56]. Decision support systems that provides predictive hyperglycaemia alerts is important such that it allows individuals with diabetes, particularly those on MDI treatment regimens, to provide pre-emptively provide insulin.

These decision support systems for these functions have typically been based on algorithms that rely on a selection of heuristics (rules-based algorithms) or physiological models (physiologic model-based algorithm). For physiologic model-based algorithms, a number of physiological models of the glucose-insulin dynamics [57, 58, 59] form the basis of these algorithms. These models also describe the effect of subcutaneous insulin transport from insulin infusion and the rate of appearance of carbohydrates from meals. The rule-based or heuristic algorithms typically are based on a number of knowledge-based rules or fuzzy logic [60, 61].

The success of data-driven models, particularly deep learning, in areas such as vision and natural language processing along with the burgeoning promise of data-driven models in the field of healthcare has led to a renewed focus on the application of current machine learning techniques for developing the next generation of decision support systems.

2.4 Machine Learning for Diabetes Technology

In the process of managing diabetes, activities ranging from daily glucose control to scheduled clinical tests generate large amounts of data that can provide further insight on analysis. In this section we establish the formulation of the problem for developing data-driven models as decision support systems. We also briefly cover a number of different decision support systems that have been developed in the literature and covered in this thesis.

The adoption of wearable technologies that enable continuous, minimally invasive means of monitoring users that generate large amounts of data. This enables short term predictions that can help in the daily management of diabetes with the main aim of maximising time in target. Machine learning can be used to develop models that provide meaningful alerts when the user is about to succumb to an adverse glycaemic event. The majority of research directions in this area cover supervised

learning approaches.

In supervised learning, the aim is to learn a mapping $f : X \rightarrow Y$, where $X \in \mathbb{R}^{k \times n}$ represents the input domain with k data points and n unique features. y is the output domain that comprises the associated ground truth given a value of the input (covariate), X. The task is termed a classification task when of the output is discrete/categorical i.e.$f : \mathbb{R}^k \rightarrow [1, ..., m]$ and a regression task when the output is represented as continuous values i.e. $f : \mathbb{R}^k \rightarrow \mathbb{R}$.

Blood glucose concentration prediction is currently the leading direction for the application of machine learning in diabetes management. The current thinking underpinning this problem formulation is that while detection of glucose concentration levels aids in minimising the time spent in hyperglycaemia or hypoglycaemia, prediction of such events would allow pre-emptive actions to potentially avoid them. In a machine learning framework, this problem is formulated in two main ways and tackled with the various learning algorithms.

Table 2.1: A review of different algorithms applied in the field of diabetes management

Algorithm	Subjects	Inputs	Prediction Horizon	Metrics	Reference
Glucose Level & Adverse Event Forecasting					
Linear Regression	89 real	CGM	30 min	F1, Delay	[52]
Bayesian Regression	89 real	CGM	30 min	F1, Delay	[52]
Support Vector Classifier	89 real	CGM	30 min	F1, Delay	[52]
Latent Variable Model	10 *in-silico* 7 real	CGM, Insulin, Meals	30 - 60 min	RMSE, EGA	[62]
Exogenous Autoregressive Model	9 real	CGM, Insulin, Meals	30 - 90 min	RMSE, FIT	[63]
Artificial Neural Network	9 real 6 real	CGM	15 - 45 min	RMSE, Delay	[64]
Support Vector Regression	27 real	CGM, Insulin, Meals Exercise, Time	15 - 120 min	RMSE, CG-EGA	[65]
Gaussian Process Regression	6 real	CGM, Insulin, Meals Sleep, Work, Exercise, Heart rate Skin conductance, Skin temperature	30 min	RMSE, Precision, Recall	[66]
Random Forest	27 real	CGM, Insulin, Meals Exercise, Time	15 min - 120 min	RMSE, EGA, Correlation Coefficient	[67]
Personalised Nutrition					
Boosted Decision Tree	800 real	Gut microbiome, Blood tests Questionnaires Food diary Anthropometrics	-	iAUC	[68]
Bolus Recommendation					
Case-Based Reasoning (CBR)	20 *in-silico*	Time, CGM, Meals, Exercise, Insulin features	-	Mean BG, Glycaemic Targets BGRI, CVGA	[69]
K-Nearest Neighbours (kNN)	100 *in-silico*	CGM, Insulin, Meals, Rate of Change	-	Mean BG, RMSE, TDI, Glycaemic Targets	[70]
Artificial Neural Network	100 *in-silico*	CGM, Insulin, Rate of Change	-	BGRI	[43]
LASSO	100 *in-silico*	CGM, Meals, Insulin	-	BGRI, Glycaemic Targets	[44]
Meal Detection and Estimation					
Ensemble Approach (MDA)	19 real	CGM, Meals, Insulin	-	Delay	[71]
Glucose Rate Increase Detector	10 *in-silico*	CGM, Meals, Insulin	-	False Positives, True Positives, Delay	[72]
Adaptive Unscented Kalman Filter	100 *in-silico*	CGM, Meals, Insulin, Heart rate	30 min	Precision, Recall, Delay	[50]
Variable State Dimension (Extended Kalman Filter)	30 *in-silico* 2 real	CGM, Meals, Insulin	30 min	Sensitivity, False Positive Rate, Delay, Error Size	[48]
Rule-based + Fuzzy logic	30 *in-silico*	CGM, Meals, Insulin	-	Sensitivity, False Positive Rate, Delay, Error Size	[47]
Unscented Kalman Filter	10 *in-silico*	CGM, Meals, Insulin	-	Sensitivity, False Positive Rate, Delay	[49]
Logistic Regression	100 *in-silico*	CGM, Insulin	-	Delay	[73]

As seen in Table 2.1 above, there are myriad problems that are formulated as machine learning tasks. The additional challenge lies in the fact that multiple metrics exist for evaluating the performance of the stated approaches. In the case of glucose level prediction, the general task is typically formulated as a regression problem where the label is the glucose value over a stated prediction horizon, typically 30 minutes.

For adverse glycaemic event prediction, the task is formulated as a supervised classification problem. The labels usually cover the set of possible glycaemic events over the prediction horizon: hyperglycaemia, euglycaemia, and hypoglycaemia. Gadaleta et al. [52] the levels are initially classed: Severe hypoglycaemia, hypoglycaemia, euglycaemia, hyperglycaemia, and severe hyperglycaemia. The authors note that an improvement in classifier performance is observed when specialized, in this case, to predict hyperglycaemia. Finally, while the F-measure is important it is necessary to consider the time lag as while, hence the authors finally settle on the specialised SVM classifier (F1 = 0.73) although the linear regression and bayesian regression methods have marginally better performance (F1 = 0.76).

An observation from these works reveals reduced performance as the prediction horizon increases. This can be due to the increasing influence of external factors such as future meals, physical activity and/or stress that can occur in the intervening period.

Current results from these works have also highlighted the benefit in predicting glycaemic state of subjects. Zeevi et al. [68] use glycaemic response to meals to personalize nutrition for T2DM subjects and improve postprandial glycaemic response; performance is evaluated by the difference between responses to a good diet week (19 mg/dL) and bad diet week (50 mg/dL). This approach is in contrast to other approaches that then consider modifications to bolus recommendations to improve the glycaemic outcome. Herrero et al. [69] employ case-based reasoning as means of adapting meal boluses to reduce the potential period of post-prandial

hyperglycaemia. This is considers the most relevant previous cases in order to determine the appropriate insulin bolus to recommend. Cappon et al. [43], on the other hand, modulate the standard formula for bolus recommendations using a 3-layer feedforward neural network to optimise recommendation. This has shown moderate success in reducing the BGRI relative to the standard formula.

Evidently, the current state of diabetes management tools that have been devised over the years to enable effective glucose management have been implemented to varying levels of success and with challenges remaining. As noted in this section, data-driven approaches have provided an avenue for beginning to tackle the arising challenges. In the following chapters, we introduce deep learning methods for tackling these challenges and advancing the current state of diabetes technology across the range of tools available to individuals with diabetes.

Chapter 3

Deep Learning for Blood Glucose Prediction

3.1 Introduction

The standard approach to diabetes management requires people actively taking BG measurements a handful of times throughout the day with a finger prick test - self monitoring of blood glucose. The recent development and uptake of continuous glucose monitoring (CGM) devices allow for improved sampling (5 minutes) of glucose measurements [74]. This approach has proven to be effective in controlling BG and thus improving the outcome of subjects in clinical trials [75]. Further improvement of glucose control can be realised through prediction, which allows users to take actions ahead of time in order to minimise the occurrence of adverse glycaemic events. The challenges lie in multiple factors that influence glucose variability, such as insulin variability, ingested meals, stress and other physical activities [76]. In addition, individual glycaemic responses are conditioned by high subject variability [77, 78], leading to different responses between individuals under the same conditions.

Machine learning (ML) allows intelligent systems to build appropriate models by learning and extracting patterns in data. The models discover mappings from the representation of input data to the output. Performances of traditional machine learning algorithms such as logistic regression, k-nearest neighbours [79], or support vector regression [80] heavily rely on the representation of the data they are given. Typically, the features - information the representation comprises - are engineered with prior knowledge and statistical features (mean, variance) [81], principal component analysis (PCA) [82] or linear discriminant analysis [83]. Artificial neural networks (ANN) are also investigated widely in diabetes management [84, 85, 86, 87, 88]. One advantage of the artificial neural network is that with enough layers, feature engineering may not be necessary for modelling. However, ANN models in the literature are mostly implemented with fewer than 3 layers, hence its learning capacity is limited due to the model complexity. Deep learning, which incorporates multi-layer neural networks, has lead to significant progresses in computer vision [89], diseases diagnosis [90], and healthcare [91, 92]. Deep learning shows superior performance to traditional ML techniques due to this ability to automatically learn features with higher complexity and representations [93, 94, 95, 96].

In this chapter, we propose a deep learning algorithm for glucose prediction using a multi-layer convolutional recurrent neural network (CRNN) architecture. The model is primarily trained on data comprising CGM, carbohydrate and insulin data. After preprocessing, the time-aligned multi-dimensional time series data of BG, carbohydrate and insulin (other factors also can be considered) are fed to CRNN for training. The architecture of the CRNN is composed of three parts: a multi-layer convolutional neural network that extracts the data features using convolution and pooling, followed by a recurrent neural network (RNN) layer with long short term memory (LSTM) cells and fully-connected layers. The model is trained end-to-end.

The convolutional layer comprises a 1D Gaussian kernel filter to perform the temporal convolution, and pooling layers are used for reducing the feature set. A variant of recurrent LSTM model is leveraged since LSTM shows good performances in pre-

dicting time series with long time dependencies [97]. The final output is a regression output by fully connected layers. The CRNN model is realized using the open-source software library Tensorflow [98], and it can be easily implemented to portable devices with its simplified version Tensorflow Lite. The performance of the proposed method is evaluated on datasets of simulated cases as well as clinical cases of T1DM subjects, and compared against benchmark algorithms including support vector regression (SVR) [80], the latent variable model (LVX) [99], the autoregressive model (ARX) [63], and a neural network for predicting glucose (NNPG)[84].

3.2 Datasets

In order to evaluate the performance of the proposed approach we consider two datasets: an *in-silico* dataset and a clinical dataset. This section details the two datasets employed in this work and the characteristics of both datasets.

The *in-silico* dataset consists of 10 adult T1DM subjects and was generated using the UVA/Padova T1D, which is a simulator for glucose level simulation approved by the Food and Drug Administration (FDA) [100]. This simulator serves as robust and validated framework for generating simulated cases. The cohort of T1D cases generated can be configured with varying meal and insulin information such that each case sufficiently differs. In this work, we used a modified version of the simulator which includes such variability. In particular, the variability on meal composition, insulin absorption, carbohydrate estimation and absorption, and insulin variability were included. In addition, a simple model of physical exercise was also used. Further details on how the simulator was modified can be found in [101].

The 10 unique adult cases each has 360 days of data for each case. This covers 3 meals per day - breakfast, lunch, and dinner. Insulin entries vary in each day, from 1 to 5. The insulin entry can be with a meal (meal and insulin at almost the same time), or without a meal (correction bolus). A simple exercise model is considered at certain points, which occur occasionally at any time of the day except for during

the nighttime. The training and validation set account for 50% of the dataset, and the testing set is the rest of data.

On the other hand, the clinical dataset was obtained from a clinical study at Imperial College Healthcare NHS Trust St. Mary's Hospital, London (UK) consisting of multiple phases evaluating the benefits of an advanced insulin bolus calculator for T1D subjects [102]. The dataset in consideration was collected from a 6-month period involving 10 adult subjects with T1D. The information included in the dataset comprises glucose, meal, insulin, and associated time stamps. In building the dataset, we mainly consider CGM and self-reported data such as insulin boluses and meals, as is done with the *in-silico* dataset. Before that, we excluded participants whose data exhibited large gaps (corresponding to weeks of missing data), insufficient reports of exercise over the 6-month period, and extensive errors in sensor readings.

The CGM data was measured using Dexcom G4 Platinum CGM sensors, with measurements received every 5 minutes. The CGM sensors were inserted from the first day of the study, and calibrated according to the manufacturer instructions. Other information available in the dataset such as meal, insulin, and exercise were logged by the subjects in a mobile application on a smartphone. Though the selected data has good quality, many periods of missing data, bad points or unexpected fluctuations exist relative to the quality of the *in-silico* dataset. Similar to the *in-silico* dataset, each subject's clinical data is halved for training and testing data.

The CRNN model can be applied to datasets where other inputs are available, such as self-reports of exercise, stress and alcohol consumption. We believe that these information are useful and can increase the forecast accuracy in some cases. However, in this work we only consider CGM data recorded every 5 minutes, meal data indicating meal time and amount of carbohydrates, as well as insulin data with each bolus quantity and the associated time as input in the model.

3.3 Methods

In this section we explain the proposed approach, convolutional recurrent neural network (CRNN), in more detail. The approach consists of several components: preprocessing, feature extraction using CNN, time series prediction using LSTM and a final conversion to the final output. The architecture of the proposed CRNN is shown in Fig. 3.1. In the diagram below, the input of the algorithm is time series of glycaemic data from CGM, carbohydrate and insulin information (time and amount); other related information are optional (exercises, alcohol, stress, etc.). The output of preprocessing is cleaned, time-aligned glycaemic, carbohydrate and insulin data, which are then fed to the CNN.

The output of CNN serves as the input of RNN, which is a multi-dimensional time series data, representing the concatenation of features of the original signals. The output of the RNN is the predictive BG level 30-min (or 60-min) later, while hidden states are inherited and updated continuously internally inside of the RNN component. The model is trained end-to-end. We evaluate the models with 30 and 60-min prediction horizon (PH) because it is widely used in glucose prediction software, and is easier to compare results with other works [103, 104, 86, 80, 99, 84]. We proceed with an explanation of the data pipeline and components of the model architecture.

3.3.1 Data Preprocessing

The main purpose of the preprocessing component is to clean the data, filter the unusual points and make it suitable as the input to the neural network. Besides the normal steps including time stamp alignment and normalization, the most important operation to improve the data quality is the outlier detection, interpolation/extrapolation and filtering, in particular for clinical data. Because in clinical data, there are many missing or outlier data points due to errors in calibration, measurements, and/or mistakes in data collection and transmission. Here, several methods can be used to handle these scenarios [105]. They include dimension re-

duction model to project data into lower dimensions [106], proximity-based model to determine the data by cluster or density [107], and probabilistic stochastic filters [108] to rule out outliers.

For cases when the data fluctuates with high frequency, 1D Gaussian kernel filter is implemented on the glucose time series. A smoothed continuous time series of glycaemic data is then obtained along with the time-aligned carbohydrate and insulin information. In this work, for *in silico* data we do not use filters because the dataset is already clean. For clinical data used in this work, we use the Gaussian filter. The 3-dimensional time series that covers the last 2 hours before the current time is sent to the neural network as input. A sliding window of size 24 is employed to train the model. This is because during the experiments we found that 24 is an optimal setting considering the tradeoff between the prediction accuracy and the computation complexity. In [84] the NNPG algorithm uses a similar window size of 20.

3.3.2 Convolutional Recurrent Neural Networks

Convolutional Layers

The filtered time series signal goes through the multi-layer convolutions, which transform the input data into a set of feature vectors. The convolution operation follows the temporal convolution definition shown below:

$$z[m] = \sum_{i=-l}^{l} x[i] \cdot \delta[m-i], \tag{3.1}$$

where x represents the input signal, δ denotes the kernel, z is the result of the convolution, and m is the result's index. Specifically in the first layer, x' length is the sliding window size of 24, kernel δ has a size of 8. The input signal can be fed using a sliding window setting. The windows can be overlapped or non-overlapped, determined by the allowed CNN size and computations. The convolutional layers were selected as a component in order to automatically learn the associated weights

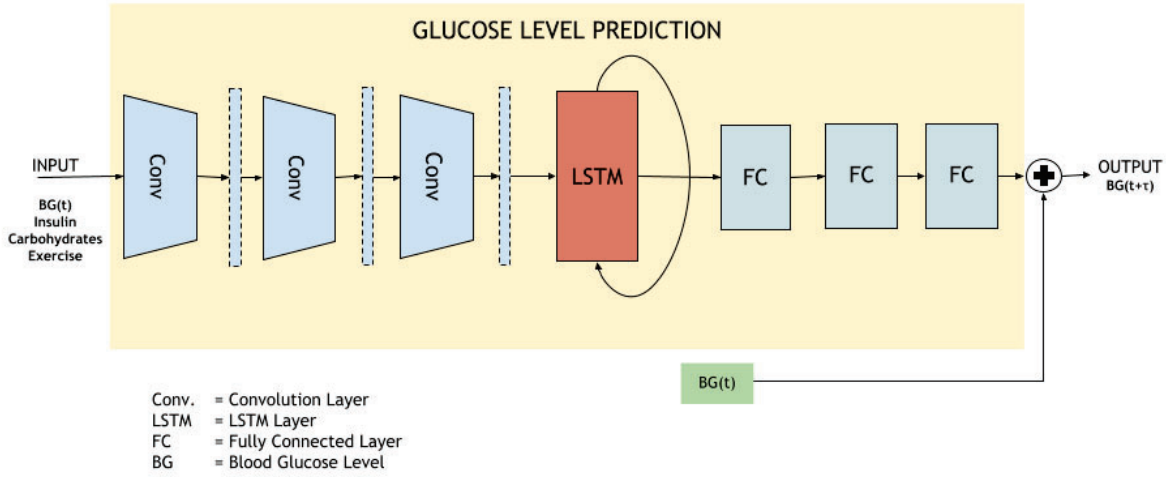

Figure 3.1: The architecture of the proposed convolutional recurrent neural network for BG prediction. The data at the left is the concatenated time series data including glucose level, carbohydrate, insulin and other factors. After outlier filtering, the multi-dimensional data can be sent to the multi-layer convolutional component. Then the resultant time series is sent to the modified recurrent neural network component presented in a red frame, which includes LSTM cells and dense fully connected layer. Finally, the resultant is converted back from "change of the glucose value" to "absolute glucose value". The output is the future glucose values of PH (eg. PH = 30 mins).

and recognises particular patterns and features in the input signal that can best represent the data for future time steps. We posit that the convolution process with the Gaussian kernel could aid in transforming the meal and insulin inputs to rates of appearance of both glucose and insulin respectively. The dimension of data in each layer is detailed in Appendix B.

The proposed method has 3 convolutional layers, with max pooling applied to down-sample the feature map obtained from the previous convolutional layer. It is common to periodically insert a pooling layer in-between successive convolutional layers to progressively reduce the size of the representation, as well as the computation. It also guards against the problem of overfitting. For instance, if the accepted size is $L1 \times D1$, and the down-sampled parameters are spatial extent F and stride S, and it results in a max-pooling vector Y of size $L2 \times D2$ as shown in Equation 3.2.

$$
\begin{aligned}
L2 &= (L1 - F)/S + 1; \\
D2 &= D1; \\
Y_i &= \max(y_i^*)
\end{aligned}
\tag{3.2}
$$

where y_i^* is the vectors after being down-sampled, Y_i is the feature map and max() is the operator that computes the maximum value. The last convolutional layer feeds directly into the recurrent layer that makes up the next component in the architecture.

Recurrent Layers

An LSTM network comprised of 64 LSTM cells is adopted as recurrent layers [109]. Each LSTM cell consists of an input gate, an output gate and a forget gate. Each of the three gates can be thought of as a neuron, and each gate achieves a particular function in the cell. The LSTM network is good at building predictive models for time series and sequential data [110]. These cells retain previous data patterns over arbitrary time intervals, thus the internal "memory" can predict the future output according to the previous states. Its memory can be updated simultaneously when new data are fed to the model. Equation 3.3 details the equations governing the internal cell state. The LSTM was selected because this architecture provided the best performance relative to other recurrent cells i.e. vanilla RNN and GRU cells.

The output of the CNN, a multi-dimensional time series, is connected to the LSTM network. We implemented an RNN with 1 hidden layer, consisting of a wide LSTM layer consisting of 64 cells. Dropout is also applied after the LSTM layer. Dropout refers to ignoring neurons randomly during the training phase. It has been verified that in many cases that dropout can effectively minimise issues of overfitting and improve model generalisation [111].

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
g_t &= \sigma_t(W_g x_t + U_g h_{t-1} + b_g) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_t(g_t) \\
h_t &= o_t \circ \sigma_t(c_t),
\end{aligned}
\tag{3.3}
$$

where x, h, f, i, o, c represent the input, output, forget gate, input gate, output gate, and memory cell respectively. The σ_g and σ_t, is the sigmoid activation function and hyperbolic tangent activation. and entrywise product, respectively. W, U, b represent the weights and biases that are learned during training.

The output of the LSTM which is the final state, h_t, feeds a multi-layer fully connected network, which consists of 2 hidden layers (256 neurons and 32 neurons) and an output layer with the glucose change as output, y_t. The fully connected layer produces the output with an activation function

$$
y_t = act(\sum_{i=1}^{N} h_t w_i + b_i) + x_t, \tag{3.4}
$$

where y_t is the multi-dimensional output, $act()$ is an activation function, w_i and b_i are weights of the fully connected network. Particularly, $act()$ can be chosen from a set of activation functions such as sigmoid function $act(a) = 1/(1 + e^{-a})$, rectifier $act(a) = \log(1 + \exp(a))$ or simple linearly $act(a) = a$. In this paper we choose the linear function $act(a) = a$ as the activation function for its simplicity.

In the training phase, a gradient descent optimisation is used. The initial weights of the network are set randomly, and the mean absolute error (MAE), shown in Equation 3.5, is set as the loss function to be minimised in the training. Partial derivatives of the error in terms of the weights w_i and bias b_i are computed, and the associated w_i and b_i are updated accordingly. The mean absolute error between

the target and the predictive value is being minimised. The optimiser we use is *RMSprop*, because it is a good choice for recurrent neural networks. It usually maintains a moving (discounted) average of the square of gradients [112], and divides gradient by the root of this average.

$$\mathcal{L}_{MAE}(y(k), \hat{y}(k|k-PH)) = \frac{1}{N}\sum_{k=1}^{N}|y(k) - \hat{y}(k|k-PH)|, \tag{3.5}$$

where $\hat{y}(k|k-PH)$ denotes the prediction results provided the historical data and y denotes the reference glucose measurement, and N refers to the data size.

The values of the hyperparameters that provide the best performance are determined by grid search. Finally, the effectiveness of the various components of the final architecture is demonstrated through a sensitivity analysis.

3.3.3 Software and Hardware

After the model has successfully undergone training and validation, we implement our algorithm on a smartphone through Tensorflow Lite due to its efficiency running on portable devices. The model is converted to a Lite model file and installed on an Android or iOS system. It needs the associated application programming interface (API) and interpreter to carry out the inference. Fig. 3.2 illustrates the manner in which the Tensorflow Lite model file [98] is wrapped and loaded in a mobile-friendly format.

Besides the CGM sensors and smartphones used in data acquisition, the proposed model is developed with Python 3.6 and Keras v2.2.2 (Tensorflow backend) [113], and trained using a NVIDIA GeForce GTX 1080 Ti.

3.4 Results

In this section we test the proposed CRNN algorithm for glucose level forecasting using a *in silico* dataset and a clinical dataset. The performance of the proposed

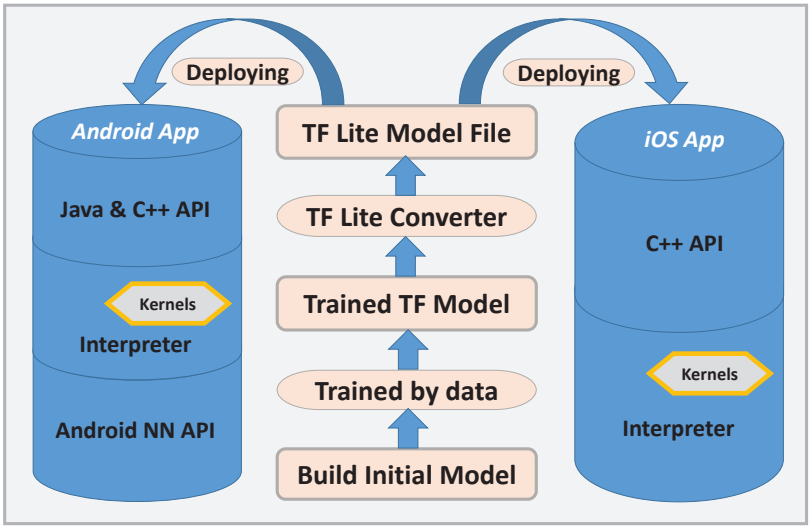

Figure 3.2: A flow diagram explains the procedure to implement deep learning models as Tensorflow Lite files to portable devices. In the figure, the yellow square frames denote the models or files obtained after each operation, and the yellow elliptic frames denote the associated operations applied. After the TF Lite model files are created, they can be deployed in Android or iOS app with slightly different settings, as shown in the cylinders on left and right.

algorithm is contrasted with that of four baseline methods: NNPG, SVR, LVX and ARX (3rd order). The results are compared with the same input data after the same pre-processing. The performance of the methods are compared based on the accuracy over 30- and 60-minute prediction horizon. In addition, we evaluate the time lag of the prediction. Different algorithms were tested on the *in silico* data generated in a way described previously. The parameters involved in these algorithms are tuned carefully for optimal result. In SVR, the SVR function in Python is applied with the optimal parameters ($C = 1e2, \gamma = 0.01, cache_{size} = 1000$). The LVX method is applied based on the MATLAB code provided in [99], the optimal predictor length and the number of LVs are $J_x = 4$ and $N_{LV} = 4$ respectively. This represents 20-minute historical data of glucose measurement, insulin and meal information being used for prediction. The 3rd order ARX model is optimized by MATLAB function $arx()$ for every specific subject.

3.4.1 Criteria for Assessment

Several criteria are used to test the performance of the proposed algorithm. The root-mean-square error (RMSE) and mean absolute relative difference (MARD) between the predicted and reference glucose readings serve as the primary indicators to evaluate the predictive accuracy.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y(k) - \hat{y}(k|k-PH))^2}, \tag{3.6}$$

$$MARD = \frac{1}{N}\sum_{k=1}^{N}\frac{|\hat{y}_k(k|k-PH) - y(k)|}{y(k)}, \tag{3.7}$$

The RMSE and MARD provide an overall indication of the predictive performance. As mentioned earlier, the benefit of glucose prediction is avoiding adverse glycaemic events. In the clinical context, these metrics are limited in the insight they provide. Additional metrics are needed to assess the proposed algorithm in the following perspective:

- Capability of the forecasting algorithm in differentiating between adverse glycaemic events and non-adverse glycaemic events.
- Time delay in the predicted glucose readings and reference values to evaluate the response time provided to deal with the potential adverse glycaemic event.

The Matthews Correlation Coefficient (MCC) is used to evaluate the performance of the algorithms for detecting either adverse glycaemic event (hypoglycaemia or hyperglycaemia).

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{3.8}$$

where TP, FP, FN, TN stand for true positive, false positive, false negative, and true negative respectively. In this case, a true positive indicates a correct classification of

hypoglycaemia (< 70 mg/dL) or hyperglycaemia (> 180 mg/dL) event in the next 30 or 60 minutes. We consider a true adverse event to have occurred when either scenario persists in the CGM data for at least 20 minutes [114]. In addition, we consider an event a true positive when the predicted event is at most 10 minutes (PH+2 timesteps ahead) leading or within 25 minutes of the prediction horizon lapsing (1 timestep for PH = 30 min, and 7 timesteps for PH = 60 min) the reference event.

A standard confusion matrix typically includes the Accuracy as opposed to Matthews Correlation Coefficient (MCC). This modification addresses the imbalance in classes inherent in this situation - non-adverse events far outweigh adverse events.

The effective prediction horizon is defined as the prediction horizon, taking into account delays due to the responsiveness of the algorithm for a predicted value relative to its reference value. Cross correlation of the predicted and actual readings is employed in performing a time delay analysis of the proposed algorithm to determine the effective prediction horizon.

$$\begin{aligned} PH_{eff} &= PH - \tau_{delay} \\ &= PH - \arg\max_{\tau}(\hat{y}_k(k|k-PH) \star y(k)). \end{aligned} \tag{3.9}$$

A singular quantitative metric is not sufficient in evaluating performance of the proposed algorithm. Consequently, the set of metrics collectively give a comprehensive description of the quality of the prediction algorithm performance.

p-values are calculated for other algorithms comparing to the proposed algorithm in terms of smaller RMSE, MARD or longer PH_{eff}. A Shapiro-Wilk Test is used to ascertain the normality of the results before performing a paired t-test to derive the p-values. Across all results, the tests show the null hypothesis (samples drawn from a Gaussian distribution) cannot be rejected.

3.4.2 Performance comparison with *in-silico* data

Table 3.1: Comparison of the prediction accuracy and of different prediction methods for 10 virtual adult diabetic subjects (best result highlighted in **BOLD**)

PH (min)	Metric	CRNN	NNPG	SVR	LVX	ARX
				Overall		
	RMSE	**9.38±0.71**	12.91±1.19‡	12.48±1.94‡	11.32±1.34†	13.27±1.19‡
	MARD	**5.50±0.62**	7.05±0.94‡	6.40±1.36	6.59±0.80‡	7.46±1.02‡
30	PH_{eff}	**29.0± 0.7**	20.8 ± 1.8‡	23.3 ± 1.6‡	27.5 ± 1.3†	20.5 ± 1.7‡
				Hyperglycaemia		
	MCC	**0.84±0.05**	0.83±0.05	0.84±0.05	0.83±0.05	0.81±0.05
				Hypoglycaemia		
	MCC	0.79±0.15	0.64±0.20	0.79±0.10	**0.83±0.06**	0.78±0.10
				Overall		
	RMSE	**18.87±2.25**	24.24 ± 3.01‡	23.46 ± 3.33‡	22.42 ± 2.74‡	25.73 ± 3.24‡
	MARD	**9.16±1.16**	13.70 ± 1.88‡	10.83 ± 1.48	12.20 ± 1.82‡	13.75 ± 2.45‡
60	PH_{eff}	**49.8±2.9**	31.0 ±4.7‡	32.6 ± 4.1‡	44.2 ± 2.7‡	19.8 ± 2.7‡
				Hyperglycaemia		
	MCC	0.82 ± 0.05	0.79 ± 0.06	0.78 ± 0.07	**0.86 ± 0.04**	0.64 ± 0.05
				Hypoglycaemia		
	MCC	**0.80 ± 0.14**	0.38 ± 0.39	0.79 ± 0.10	**0.80 ± 0.07**	0.72 ± 0.12

$^{*}p$-value < 0.05;$^{\dagger}p$-value < 0.01;$^{\ddagger}p$-value < 0.005

The results of RMSE, MARD and forecasting of adverse glycaemic events are summarized in the Table 3.1. In the Table we compare the predictive error of the algorithms to measure the accuracy of the algorithms. The CRNN algorithm exhibits the best overall RMSE and MARD for the 10 simulated cases at short(30) and long term (60) predictions. The results in Table 3.1 are statistically significant relative to each algorithm. This observation is also evident in both the hyperglycaemia and hypoglycaemia region. In the hyperglycaemia region the CRNN shows a statistically significant improvement in the glucose prediction over most other algorithms, with the exception of LVX. CRNN reports a statistically significant improvement in effective prediction time (+1.5 min for 30-min and +5.6 min for 60-min) over LVX. An overall improved prediction time . On the whole, the CRNN can be evaluated as the best algorithm. The CRNN model also reports relatively low standard deviations from which we infer a benefit in building individualized models.

An illustration of a comparison of various algorithms for 30-minute shown in Figure 3.4 for a virtual adult 4. As seen in the Figure 3.4, the CRNN exhibits the best

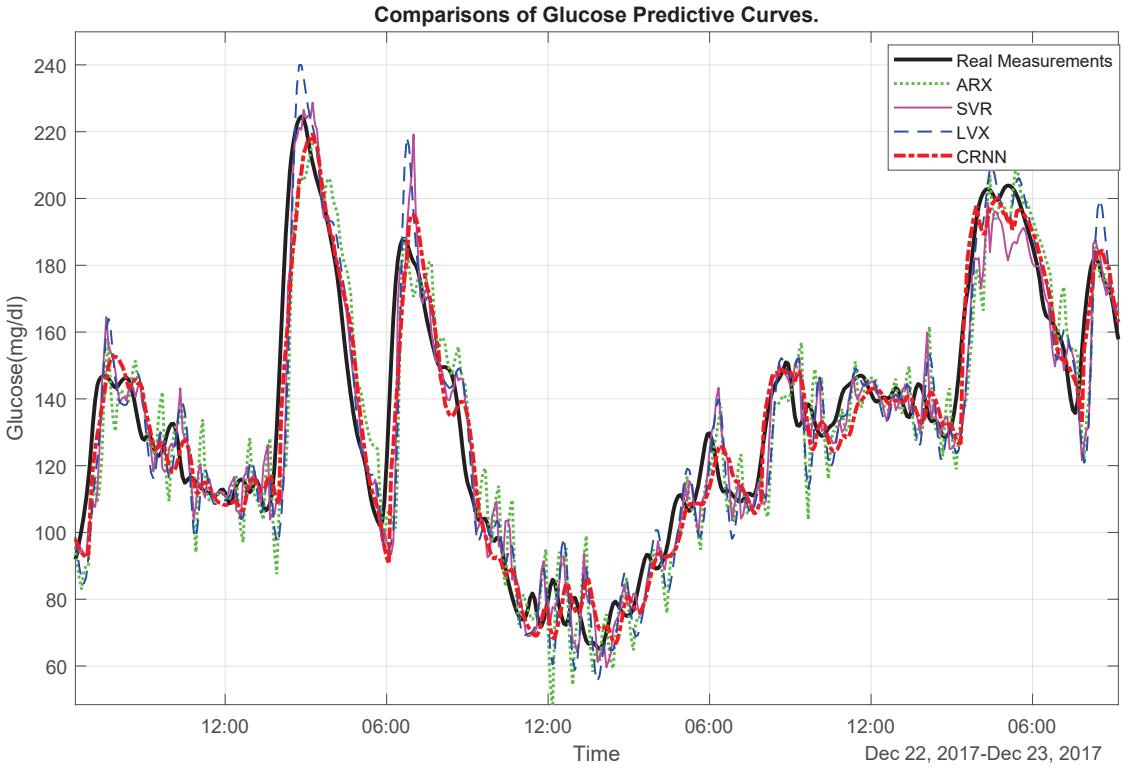

Figure 3.3: One-day period prediction results for virtual adult 4. The solid black line, dotted green line, solid magenta line, dashed blue line, dash-dotted red line indicate the simulated glucose measurements, the prediction results of the 3rd order ARX method, the prediction results of the SVR method, the prediction results of the LVX algorithm, the prediction results of the CRNN method, respectively.

responsiveness as the predictive curve responds rapidly towards the sharp glycaemic uptrend. The algorithm learns representations that appropriately account for both sharp slopes and gradual increments in the glycaemic curve. Consequently, at a glycaemic peak, CRNN yields a predictive curve with even higher slope to compensate the time lag aiming at reducing the gap between the prediction and real measurements. This feature helps CRNN to decrease the RMSE and MARD as well as maximising the effective prediction horizon.

3.4.3 Performance comparison with clinical data

As mentioned in the previous section, the data obtained in the clinical trial exhibits missing data, and erroneous data. This results in non-physiological discontinuities that would affect the training process. To mitigate these occurrences, the data

is processed with interpolations/extrapolations for gaps in data. The interpolation/extrapolations points are not included in the evaluation of the performance of the methods.

Table 3.2: Comparison of the performance metrics of different prediction methods for 10 clinical adult diabetic subjects (best result highlighted in **BOLD**)

PH (min)	Metric	CRNN	NNPG	SVR	LVX	ARX
			Overall			
	RMSE	**21.07 ± 2.35**	23.14 ± 2.99	22.00 ± 2.83	21.51 ± 2.44	21.56 ± 2.53
	MARD	11.61 ± 2.18	13.42 ± 2.35	13.54 ± 2.88	**10.93 ± 1.87**	11.00 ± 1.81
30	PH_{eff}	**19.3 ± 3.1**	12.8 ± 5.9	18.6 ± 2.8	14.5 ± 3.4	12.0 ± 3.0
			Hyperglycaemia			
	MCC	0.79 ± 0.04	0.75 ± 0.04	0.79 ± 0.05	0.79 ± 0.04	0.77 ± 0.04
			Hypoglycaemia			
	MCC	0.51 ± 0.20	0.12 ± 0.12‡	0.11 ± 0.08‡	**0.55 ± 0.17**	0.53 ± 0.15
			Overall			
	RMSE	**33.27 ± 4.79**	36.05 ± 4.85‡	34.35 ± 4.55†	37.46 ± 5.04‡	36.97 ± 4.75‡
	MARD	**19.01 ± 4.46**	21.98 ± 4.87‡	20.65 ± 3.92	19.69 ± 3.70‡	19.65 ± 3.55‡
60	PH_{eff}	**29.3 ± 9.4**	18.3 ± 4.9‡	28.4 ± 5.2	19.9 ± 5.1*	14.6 ± 5.6‡
			Hyperglycaemia			
	MCC	0.72 ± 0.05	0.66 ± 0.09*	0.74 ± 0.07	**0.76 ± 0.05**	0.71 ± 0.05
			Hypoglycaemia			
	MCC	0.40 ± 0.13	0.01 ± 0.00‡	0.06 ± 0.08‡	**0.56 ± 0.14***	0.51 ± 0.15

$^{*}p$-value < 0.05; $^{\dagger}p$-value < 0.01; $^{\ddagger}p$-value < 0.005

Table 3.2 shows the RMSE and MARD of the performance of the algorithms for the 10 cases of real data. Contrary to the relative performance of the methods in the in-silico dataset, the evaluation of the methods is mixed. The CRNN maintains the best results for RMSE and MARD over a 30 minute prediction horizon baseline methods. However, the ARX and LVX models show improved performance in terms of MARD relative to the CRNN. In addition, the LVX shows marginally better performance over CRNN in predicting adverse glycaemic events. The time delay in predicting these results shows that CRNN exhibits the best performance with the smallest lag.

Over a long-term prediction horizon, the CRNN provides the best performance for prediction of glucose level and with the least lag of the evaluated methods. SVR is able to perform close to the CRNN in terms of effective prediction horizon. However, the better prediction of hyperglycaemic events is contrasted with very poor

prediction of hypoglycaemia. The results for hypoglycaemia prediction, equivalent to random guessing, suggest that 60 minutes represents the limit of meaningful hypoglycaemia prediction for SVR and NNPG given these inputs. Further improvement may require the inclusion of engineered features. Although LVX exhibits superior performance in predicting adverse glycaemic events, it should be noted that the user would have considerably less time (-9 mins) to take action. As seen in Figure 2.5, the CRNN and LVX both achieve good predictive curves compared to the ground truth measurements. Specifically, at the inflection periods during peaks and troughs, the LVX tends to have higher and lower predictions, respectively. The CRNN follows the trend at both local and global peak points closely, which increases its overall accuracy.

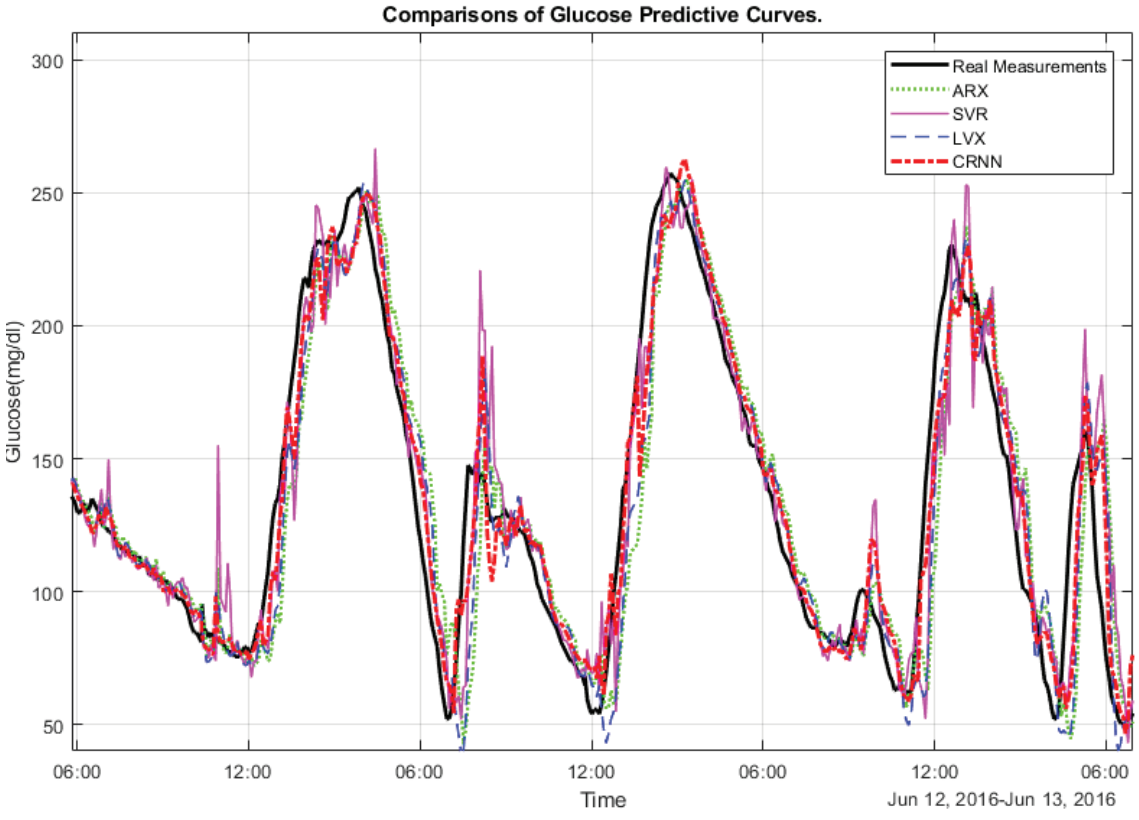

Figure 3.4: One-day period prediction results for clinical adult 17. The solid black line, dotted green line, solid magenta line, dashed blue line, dash-dotted red line indicate the real glucose measurements, the prediction results of ARX, SVR, LVX, and the CRNN algorithm, respectively.

To understand the effect of each network component, we generate networks with

different components and evaluate their performances. The results are shown in Table 3.3. It shows that full CRNN achieves the best performance, and both CNN and LSTM component contribute to the final result. In addition, we investigate the influence of different lengths of training set. The results are shown in Table 3.4. Using 1 month training data, the RMSE of CRNN can achieve 22.28 ± 2.67 (30) and 35.56 ± 4.55 (60). This can be slightly improved if longer training data are exploited. It shows that collecting more training data can increase the predictive accuracy.

Table 3.3: An ablation study showing the effect of each stage on CRNN performance

Model	RMSE	
	30 min	60 min
CRNN	21.07 ± 2.35	33.27 ± 4.79
CRNN w/o CNN	22.16 ± 4.39	36.28 ± 7.14
CRNN w/o LSTM	21.50 ± 2.62	36.01 ± 6.41

Table 3.4: A table showing the performance with different periods of training data

Training Data	RMSE	
	30 min	60 min
3 months	21.07 ± 2.35	33.27 ± 4.79
2 months	22.07 ± 2.84	35.12 ± 4.69
1 month	22.28 ± 2.67	35.56 ± 4.55

3.5 Discussion

3.5.1 Evaluation of model performance with *in-silico* and clinical data

As we establish from our results, the CRNN provides the best predictive and temporal performance compared to baseline methods.

One of the prominent improvements in the CRNN performance over the other baseline models is the effective prediction time. This significant improvement can also be attributed to the capability of deep learning to learn optimal and relevant features from the inputs for the prediction task. This can be seen in Figures 3.3. and 3.4,

where relative to the baseline methods, the response of the model to postprandial glucose increases is faster, along with faster responses to subsequent drops due to the effect of exogenous insulin.

Although the statistically significant improvements are small, the behaviour of the signals as visualised in Figure 3.3 and 3.4 offer insight into possible benefits of CRNN in clinical applications. One application with long-term glucose predictions ($\geq$ 60 min) is insulin dosing decision support [115, 42]. In this application, the predicted glucose value, particularly at local maxima and minima could be important in estimating the optimal insulin dose.

Subsequently, the evaluated models are observed to exhibit different behaviour relevant to this application. The CRNN models tend to smoothly follow the reference glucose profile with minimal lag. On the other hand, the LVX model tends to overshoot at local maxima and undershoot at local minima which can, for example, lead to erroneous estimations of any resulting insulin bolus. Similarly, the SVR and ARX models also show oscillatory behaviour and are therefore likely to also lead to more erroneous estimations.

In the previous section we noted a discrepancy in the performance of the proposed algorithm and baseline algorithms in simulated cases and the real patient cases. Previous tests have also indicated that the performance in real subjects is much less satisfactory in comparison to virtual subjects. In our opinion, the drop in performance can be primarily attributed to the increased complexity of real data generated from a patient relative to the simulated data generated from a physiological model. In addition, the gaps in data and subsequent method of interpolation/extrapolation may contribute to the further reduction in performance.

Relative to the baseline algorithms, the CRNN is better at capturing the features since deep learning affords a better capacity at learning optimal representations of features. This could also explain the relatively lower variance in metrics for the performance of the CRNN in different cases relative to baseline models.

3.5.2 Comparison with results in the literature

We achieved a mean RMSE = 9.38mg/dL *in silico* using the proposed method, and it is the best amongst other algorithms, including SVR, LVX and 3rd order ARX. In addition, we want to compare our algorithm with other approaches in the literature. Using the dataset generated from the simulator, our algorithm is better than the results of RMSE = 18.78mg/dL [103] and RMSE = 13.65mg/dL in [104]. For several other works, it is difficult to evaluate the RMSE through direct comparison due to the unavailability of the original code, model parameters, and the benchmark datasets. However, we may compare the results with widely used methods as benchmarks, such as SVR or NNPG. For instance, for PH = 30 min as shown in Table 3 [86], the algorithm is 0.1 mg/dL better than the result of SVR in terms of RMSE on the real dataset; our algorithm is 0.9 mg/dL better than the SVR in terms of RMSE on the real dataset. In [85], for PH = 30 min their RMSEs are 1.3 mg/dL better than NNPG for the simulated data and 0.2 mg/dL better than NNPG for the real datasets. Our RMSEs are 3.5 mg/dL better than NNPG for the simulated data and 2.1 mg/dL better for the real datasets. As far as we know, the proposed algorithm achieves a performance state-of-the-art accuracy with regard to RMSE. To build a fair comparison, we provide all benchmark models the same input, including CGM data, meal and insulin. For the conventional NNPG, it only uses CGM measurements. Thus, in the comparison we incorporate meals and insulin in the input as well to generate an enhanced NNPG.

3.5.3 Application on resource-constrained mobile platforms

CRNN is a personalised algorithm for different diabetic subjects. Firstly, it is data driven and personalised. Secondly, the model can be continuously updated as more data is available. In details, the model is saved as a trained neural network. We use the sequential model with Tensorflow backend to train the neural network, and the result can be saved as a small file. This file can be compiled as a ".tflite" or a ".pb"

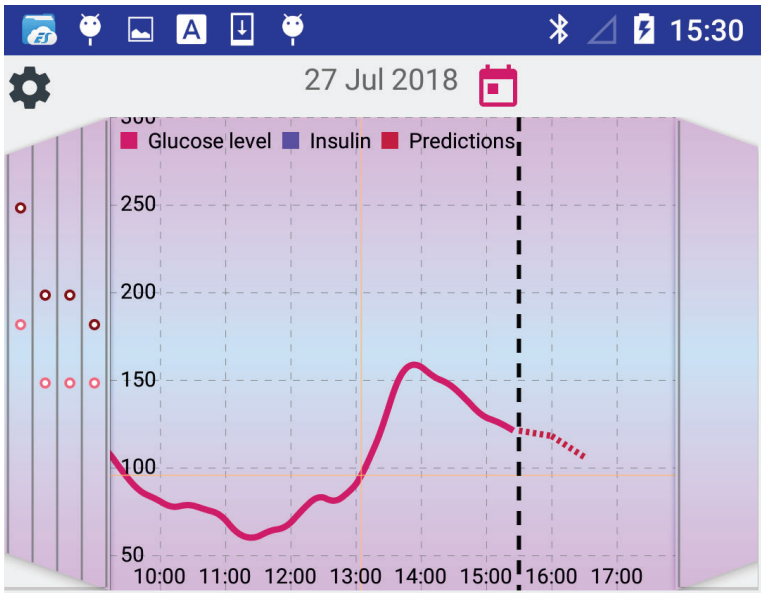

Figure 3.5: An illustration of the glucose level shown in an app interface on an Android system, where the red curve is the historic blood glucose, black dash line is the current time, and the red dot curve is the prediction provided by the model.

file for the app on mobiles, by using a Tensorflow Lite converter. The model file can be updated continuously at the cloud. The app may demonstrate the predictive glycaemic curve on screen. A demonstration on the Android system is shown in Fig. 3.5 In addition, we also found that the execution time of the model is 6ms on a Android phone (LG Nexus5 with Processor: 2.26GHz quad-core, RAM:2GB) and 780ms on a laptop (MacPro with Processor: 3.1GHz Intel Core i5, RAM:8GB). The reasons might be in the quantisation of weights and biases (e.g. 8 bit integer vs. 32 bit floating point), thus leading to a faster computation.

3.6 Conclusion

In this chapter a convolutional recurrent neural network is proposed as an effective method for BG prediction. The architecture includes a multi-layer CNN followed by a modified RNN, where the CNN could capture the features or patterns of the multi-dimensional time series. The modified RNN is capable of analyzing the previous sequential data and providing the predictive BG. The method trains models for each diabetic subject using their own data. After obtaining the trained neural network,

it could be applied locally or on portable devices. The proposed CRNN method showed superior performance in forecasting BG levels (RMSE and MARD) in the *in silico* and clinical experiments.

Chapter 4

Multitask Learning for Personalised Glucose Prediction

4.1 Introduction

In Chapter 3, we introduced a deep learning method that was a competitive approach compared to traditional machine learning approaches. The results attained are further buttressed by recent works in the literature that demonstrate deep learning [116] performing better relative to traditional machine learning approaches that usually rely on feature engineering [33, 117]. As stated previously, one advantage of the deep learning approach is that optimal features can be learned to develop personalised models for each individual [117]. However, deep learning models typically require large amounts of data in order to achieve accurate performance [118, 119].

A common challenge in the area of glucose prediction is that large amounts of subject-specific data are expensive and difficult to collect. This hampers the earlier deployment of personalised models which would aid in providing necessary inter-

ventions earlier to improve glucose control. Leveraging population data in existing datasets to address this scarcity is the complicated by inter-individual variability. Studies have suggested that differences exist in the complex glucose dynamics of individuals and clustered groups. As a result, there is a possibility that the performance of personalised models can be hampered when consideration is not given to the prior background information such as glycaemic variability [120].

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

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

We now compare this approach to inductive transfer against the popular sequential transfer learning (TL) approach that involves fine-tuning, and single-task learning (STL) models trained solely on subject-specific data. We also include support vector regression (SVR) as an additional traditional baseline approach for comparison.

4.2 Related Work

Transfer learning is a field that considers leveraging knowledge from previous experience to improve learning for a related particular task. Formally, a task $\mathcal{T}$ comprises a label space $\mathcal{Y}$ and a predictive function f that is learnt from available data. The available data is associated with a domain $\mathcal{D}$, a space that comprises the input feature space X and the output Y [121]. Each task and domain associated with previous

experience is termed a source domain and source task $\{(\mathcal{D}_S,\mathcal{T}_S)\}$ and the particular task and domain of interest is termed target domain and target task,$\{(\mathcal{D}_T,\mathcal{T}_T)\}$. Therefore in transfer learning, $\{(\mathcal{D}_S,\mathcal{T}_S)\}$ is leveraged along with $\{(\mathcal{D}_T,\mathcal{T}_T)\}$ to improve the ability to learn the target predictive function f_T.

The success of transfer learning has been noted in the fields of computer vision, digital imaging, natural language processing, as well as other areas of healthcare [122, 123, 124, 121, 125]. This typically considers using a well sourced large dataset (eg. ImageNet) in order to pre-train models, before subsequently fine-tuning models on data from the target task [118]. In order to be successful, the tasks are usually assumed to be related in some sense.

In the field of blood glucose prediction, transfer learning is not a well studied approach despite the success of data-driven methods in glucose prediction [33]. This could be attributed to the lack of large publicly available datasets in the field. For most studies that have employed deep learning, results have provided neither a consensus nor a detailed analysis on the methods and benefits of information transfer from source tasks to target tasks.

For short-term predictions (PH $\leq$ 60 minutes), the results on transfer learning are mixed. Bhimireddy *et al.* [126] employ a sequence-to-sequence network and do not report improved performance with transfer learning. On the other hand, Rubin-Falcone *et al.* [127] pre-train their models in a two-stage process; the authors first train on a large private dataset of 100 subjects before subsequently training the general model on the six subjects in the first cohort of OhioT1DM subjects and then fine-tuning on the latter cohort. This yields a 4% improvement in average predictive performance in terms of root mean square error (RMSE).

Further works have also looked at variant transfer learning approaches to leverage population data [128, 129, 130]. Hameed *et al.*[128] use knowledge distillation [131] in order to learn models that leverage large datasets. In this case, a teacher model is used to learn an initial model and the student model is subsequently trained in the target domain and the outputs - soft prediction labels - from the teacher model.

However, this approach was unable to improve performance over the student only approach trained solely on subject-specific data. Other approaches [129, 130] aim to learn from different metabolic classifications i.e. Type 2 diabetes and Type 1 diabetes. The approach of Gu *et al.*[129] is positive, although the model was only tested on one day of glucose data and the authors are unclear on whether baseline models were trained on the same data to evaluate the benefit of the transfer approach. A domain adversarial learning approach enables learning features common to each group, thereby improving generalisation and performance [130].

Kushner *et al.* [132] study the benefit of sequential transfer learning at long-term prediction horizons beyond 60 minutes. This work shows some benefit in the average performance of the neural network model on the subjects. However, further analysis suggests subjects with high glycaemic variability did not benefit from transfer learning.

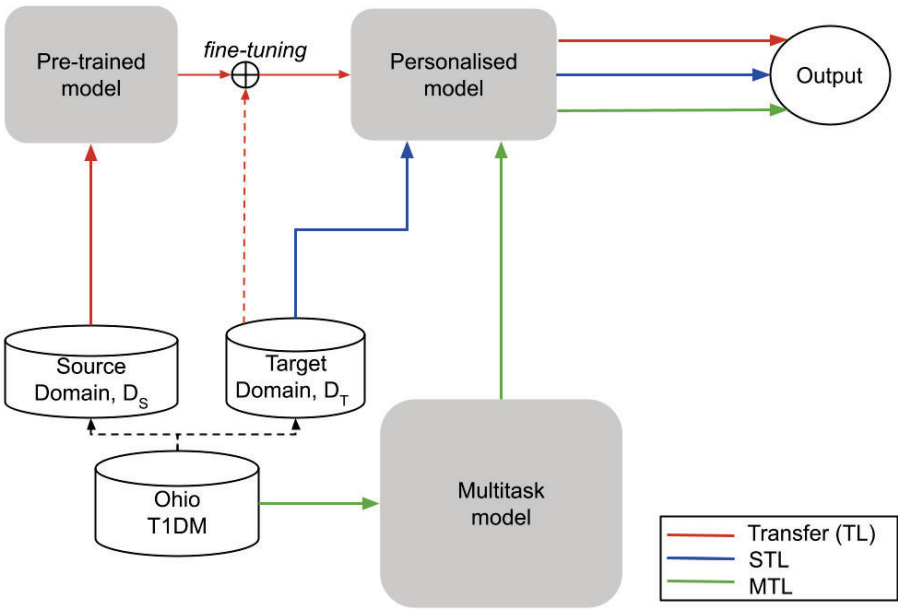

Figure 4.1: A simplified illustration of the training steps for each of the learning approaches. This covers subject-specific single-task learning (STL), transfer learning (TL) with fine-tuning, and multitask learning (MTL)

4.3 Multitask Learning

Multitask learning (MTL) is a type of transfer learning that involves learning multiple tasks simultaneously in order to improve generalisation [133]. The transfer of information signals occurs in parallel since the models for each task are jointly learned. This differs from the typical transfer learning approach which can best be described as sequential transfer, where the models are pre-trained on the data from the source domain towards source tasks before training on data from the target domain towards the target task. This approach can help to overcome the issue of inter-individual variability that exists in developing optimal personalised models within a feasible timeline.

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

4.3.1 Network Architecture

Multitask learning for neural networks is realised by sharing the parameters of the layers between the tasks in the multitask model. The architecture of the multi-task neural networks are detailed in Figure 4.2.

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

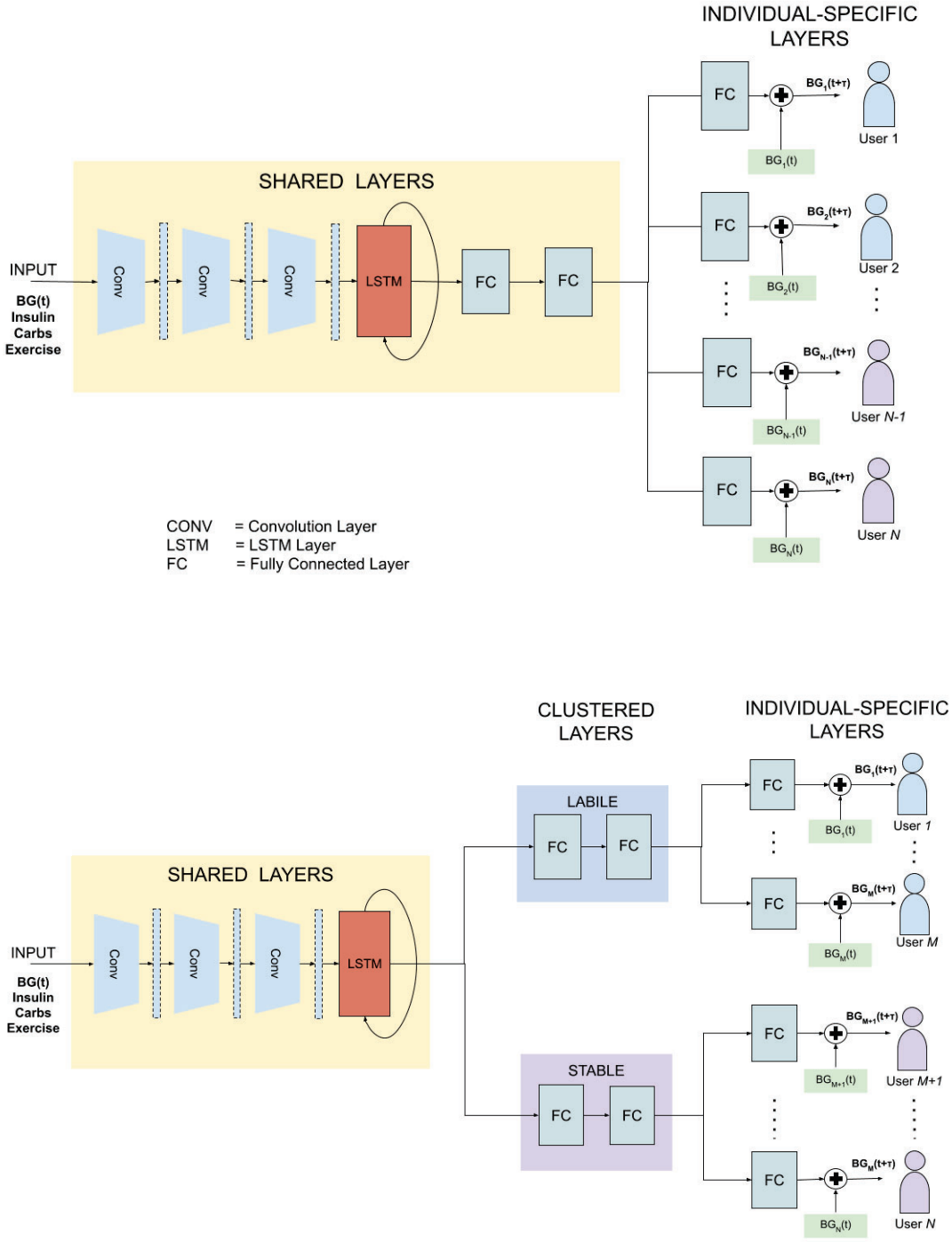

Figure 4.2: **Top:** The multitask network architecture shows the convolutional and recurrent layers are shared between all subjects. All fully connected layers except the final layers are shared by all tasks (individuals). **Bottom:** This multitask configuration shows the convolutional and recurrent layers are shared between all subjects. However, the fully connected layers are shared to varying degrees if clusters are specified.

To overcome the challenge of inter-individual variability that results in tasks being less similar, hence potential negative transfer, the degree of parameter sharing is relaxed. The intermediate layers are clustered based on prior information. For this work, we cluster individuals based on their glycaemic variability, with low (stable) glycaemic variability sharing a set of fully connected layers and high (labile) glycaemic variability .

4.3.2 Network Training

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

However, in the multitask setting, the training samples contain samples for the number of tasks (subjects) present. Consequently, samples in a mini-batch are drawn from a particular individual to train the shared layers and the layers specific to the individual similar to [124]. The loss function, $\mathcal{L}_{MAE}(y, \hat{y})$, used to minimise the error during training is defined below:

$$\mathcal{L}_{MAE}(y, \hat{y}) = \frac{1}{N_{batch}} \sum_{k=1}^{N} |y_k - \hat{y}_k|, \tag{4.1}$$

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

A sample weighting is used as a gating approach in order to ensure that the input corresponding to a subject is used to train layers in the network that pertain to the associated subject. During a forward pass the mini-batch is fed into the network to obtain the predicted glucose values, $\hat{y}$, and determine the loss according to Equation 4.1. At each iteration, the backpropagated error is used to learn personalised weights of each subject in the task specific layers and eventually learn appropriate weights

in shared layers that generalise to all subjects at the same time.

4.4 Dataset

The dataset used in this study is referred to as the OhioT1DM dataset [134], and will be referred to as such from here on. This dataset was obtained under the Data use Agreement (DUA) between Ohio University and Pennsylvania State University. The OhioT1DM dataset is updated on 2020 and comprises 12 subjects with Type 1 diabetes (T1D) monitored in free-living conditions over a period of 8 weeks.

In Chapter 3 we use an in-silico dataset and clinical dataset from the ABC4D dataset whereas in Chapter 4 we used the publicly available OhioT1DM dataset. We opted for the OhioT1DM dataset instead of the ABC4D dataset due to the OhioT1DM dataset having more individuals in the dataset. In addition to this the quality of recordings was better with fewer gaps present in CGM measurements. The adherence to logging self-reported data (such as exercise) was also better in the OhioT1DM dataset.

Table 4.1: Background information and data sizes for subjects in the OhioT1DM dataset

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

As seen in Table I above, the dataset contains 7 male subjects and 5 female subjects. In terms of age, all subjects are adults; with subjects grouped as young adults (20-40

years), middle aged adults (40-60 years), and old adults (60+ years). The glycaemic variability [120] is determined with the training data samples and is formulated through the coefficient of variation (CV) as denoted below:

$$CV = \frac{\sigma}{\mu} \times 100\%, \tag{4.2}$$

where CV is the coefficient of variation, σ is the standard deviation of the glucose concentration levels, and μ is the mean of the glucose concentration levels. A subject is classified as labile if CV $> 36\%$, and stable otherwise. This threshold represents an increased incidence of hypoglycaemia in the glucose profile and is further supported in the literature [135].

The subjects are provided with a Medtronic Enlite Continuous Glucose Monitoring (CGM) devices along with one of a Basis Peak Band (Intel Corp. Santa Clara, CA, US) or Empatica Embrace (Empatica, Inc., Boston, MA, US)[1]. The CGM measures interstitial glucose concentration levels at 5 minute intervals. The Basis Peak measures values of the skin temperature, skin conductance, heart rate, and step count (this has been aggregated over 5-minute intervals). On the other hand, the Empatica Embrace measures skin conductance, skin temperature, and acceleration magnitude (aggregated over 5-minute intervals). In addition to the physiological signals, subjects provide self-reported assessments such as meal intake, insulin, exercise, sleep, stressors, work, and sleep.

4.4.1 Preprocessing

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

In order to evaluate the models we mainly consider four features that are both prominent for the task and consistent for each subject: glucose concentration levels, insulin, meals, and exercise. The first approach is to synchronize the data entries

[1]https://www.empatica.com/

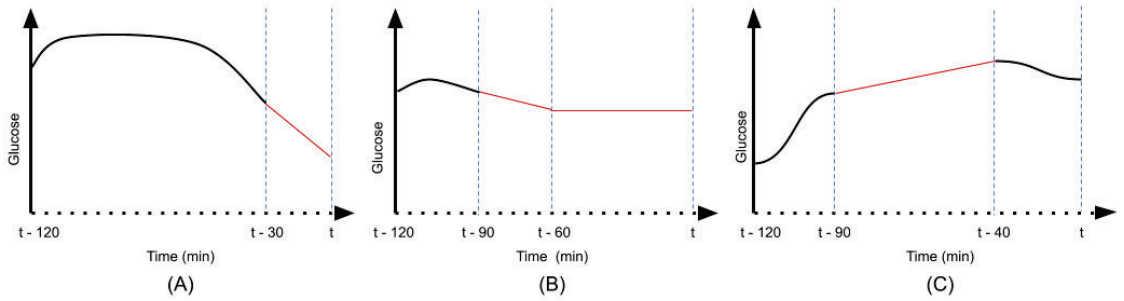

Figure 4.3: A visualisation of the imputation methods employed in this work. In (A) the 2-hour input sequence has up to 30 minutes of recent values missing and is imputed with linear extrapolation. (B) shows the imputation scheme during testing for longer than 30 minutes of recent values missing (zero-order hold). Finally (C) shows the imputation scheme when the missing values of the input sequence are located between real values (linear interpolation).

for each modality, using the CGM timestamp as the reference timestamp. A sliding window is used to extract 2-hour sequences of historical data for each prediction, selected based on analysis in previous work [136].

As noted in [134], the exercise intensity levels for each individual in the dataset are self-reported on a scale of 1-10. Given the variability between individual experiences of exercise intensity, we transform the intensity scale to a simple binary self-report on the presence (1) or absence (0) of exercise. The meals, insulin, and glucose values are scaled by 200, 100, and 120 respectively.

The non-trivial issue of missingness is next addressed in processing this dataset [137]. Missingness is present in both the physiological variables (CGM) and self-reported data and these are handled differently. In order to overcome this we implement data imputation methods in the training set and the testing set.

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

In the training set, the gaps in the glucose concentration levels are imputed using a linear interpolation. This assumes that the missing data is adequately explained

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

For accurate evaluation of the performance of the model on all test points in the test set, using only interpolation at test time is not appropriate since we may be using unknown future values. We employ different modes of extrapolation in order to impute missing CGM values in the input sequence as detailed in Figure 4.3.

4.5 Methods

4.5.1 Glucose Prediction Models

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

Support Vector Regression (SVR)

The support vector regression has been shown in the literature [65, 52] to provide competitive performance in the area of glucose prediction. This used a radial basis function (RBF) as the kernel and provides a good benchmark given the ability to perform well with small datasets. The SVR model is trained only on subject-specific data. SVR is developed using Scikit-learn v 0.21.3 Python library [138].

Deep Learning

Many deep learning architectures exist in the literature [117]. Although it is difficult to establish a standout state-of-the-art approach, most approaches are based on recurrent networks. We extend the convolutional recurrent neural network (CRNN) architecture, a 6-layer end-to-end learning framework [139], for investigating the different learning approaches detailed below. The models are developed with Python 3.6 and Keras v2.2.2 [113] and trained using a NVIDIA GTX 1050.

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

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

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

Multitask Learning (Glycaemic Variability) (MTL-GV): In this multitask setting, the training approach the same as the MTL training approach and models are trained jointly using all 12 subjects. The difference lies in the network architecture as seen in Figure 5.5.

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

The optimised hyperparameters for the various glucose models are presented in Table C.2 in the Appendix below.

4.5.2 Criteria for Assessment

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

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

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

The Clarke Error Grid Analysis (EGA) was originally developed to quantify the clinical accuracy of current blood glucose estimates against reference blood glucose values [140]. We adopt this approach, as done in literature, to evaluate the clinical accuracy of glucose forecasting algorithms as shown in Figure 4.4 [127, 132]. The graph is demarcated into five zones, labelled A-E, that represent increasing severity of errors due to misestimation of predicted glucose concentration levels as follows:

- Zone A: Predicted values in this region lie within 20% of the reference CGM values when CGM $\geq$ 70 mg/dL, and predicted CGM values are no more than 70 mg/dL during hypoglycaemia.
- Zone B: Errors of predicted values in this region fall outside the 20% error, however, any resulting standard treatment could be incorrect but uncritical.
- Zone C: Predicted values in this region could result in unnecessary treatment.

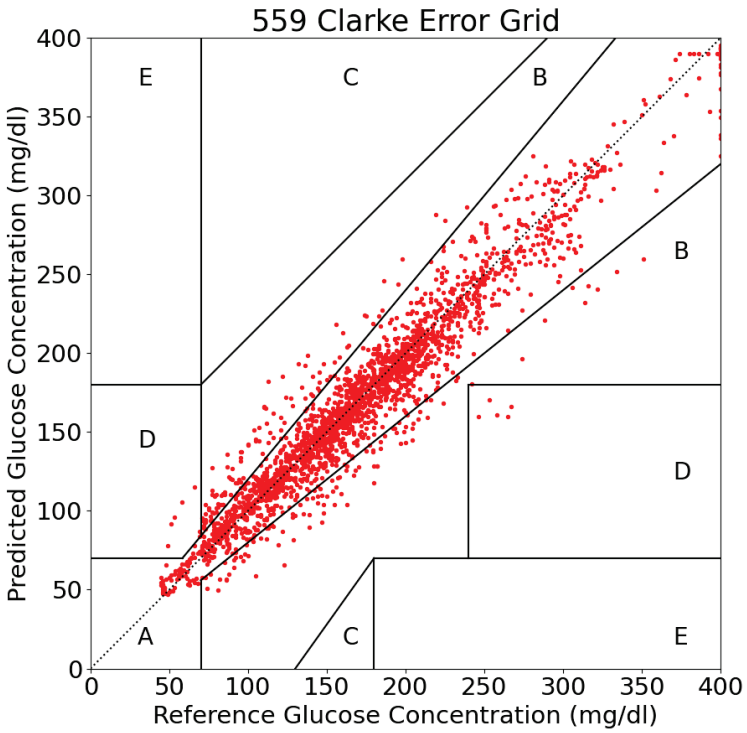

Figure 4.4: Clarke Error Grid showing the location of points in the various zones of safety for Subject 559. This illustrates the clinical relevance of errors by the MTL model at a 30-minute prediction horizon.

- Zone D: Predicted values in this region point to a potentially harmful adverse glycaemic event (hyperglycaemia or hypoglycaemia) that has gone undetected.
- Zone E: Predicted values in this region if acted on could lead to the opposite corrective action being undertaken to treat an adverse glycaemic event.

4.5.3 Statistical Analysis

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

4.6 Results

4.6.1 Performance comparison across prediction horizons

Table 4.2: Comparison of the performance metrics of deep multitask learning models against conventionally trained deep neural networks and SVR models at different prediction horizons (best result highlighted in **BOLD**)

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

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

In this experiment, we investigate the performance of the multitask learning approaches against current learning approaches and methods at prediction horizons. We evaluate the performance of MTL and MTL-GV against TL, STL, and SVR approaches for 30-120 min. MTL-GV incorporates prior information on subject glycaemic variability whereas MTL does not incorporate any prior information in the network architecture. The results for this are shown in Table 4.2.

At the short-term prediction horizon (<60 min), multitask learning shows the best performance in terms of both predictive accuracy metrics. MTL and MTL-GV showed the best mean RMSE (18.8mg/dL) and MAE (13.2mg/dL) compared to

TL (+0.4mg/dL and +0.2mg/dL), STL (+1.8mg/dL and +1.6mg/dL), and SVR (+0.4mg/dL and +0.3mg/dL) models. Compared to the conventional approach of transfer learning by finetuning (TL), MTL reveals a significant improvement (p-value $< .013$) in terms of RMSE at 30 min and in terms of RMSE and MAE at 45 min. However, the improvement in metrics compared to MTL-GV are non-significant.

The trend of MTL demonstrating the best performance remains as the prediction horizon increases. At long term predictions ($\geq$60 min), the MTL and MTL-GV models generally outperform the TL, STL and SVR models. Both MTL and MTL-GV models reveal significant improvement compared to TL models at long-term predictions in terms of both RMSE and MAE.At 120 minutes, MTL performs at least as well as the SVR models in terms of RMSE and is slightly better (-0.4mg/dL) in terms of MAE.

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

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

4.6.2 Incorporating Prior Information on Glycaemic Variability

We also examine the effect of incorporating prior information. In this study we incorporate information on glycaemic variability in the architecture of the multitask

neural network (MTL-GV).

The effect of clustering layers does not lead to an improvement in predictive accuracy until 120 min. As seen in Table 4.2, MTL consistently demonstrates the best predictive accuracy at these prediction horizons, with significant improvements over MTL-GV in terms of RMSE and MAE at 45 min and 60 min (p-value $< .05$). On the other hand, the predictive accuracy of multitask models are improved at 120 min when incorporating glycaemic variability (MTL-GV), in terms of RMSE and MAE (p-value $< .05$), over no specified clustering (MTL). The temporal gain is observed to be affected as this is lower for MTL-GV models compared to MTL models across all prediction horizons.

The Clarke Error Grid Analysis (EGA), which focuses on the percentage of samples in the safe zones (Zone A and Zone B), shows that clustering provides a slight increase in EGA (Zone A+B) at 30 min (+0.1%), 60 min (+0.3%) and 120 min (+0.1%), a decrease at 45 min (-0.4%) and no change at 90 min.

4.6.3 Impact of Training Data Size

As noted in earlier sections, most public datasets available and suitable for glucose prediction are typically small in size. This has to be considered when deploying these models in diabetes management systems. This experiment investigates the benefit of a multitask learning approach for different training set sample sizes. The prediction horizon is set at 30 minutes as it is the typical prediction horizon used in commercial CGM and predictive low glucose suspend (PLGS) systems [141].

The initial training set size covers 6 weeks. We also evaluate the performance at the following duration periods up to the end of the training set: 4 weeks, 2 weeks, 1 week and 0.5 weeks (3.5 days). Table 4.3 shows the total number of training samples for each duration period and the range of training samples for individuals. The total number of training examples in the hypoglycaemia region.

An important aspect is the performance of these models in terms of clinical accuracy

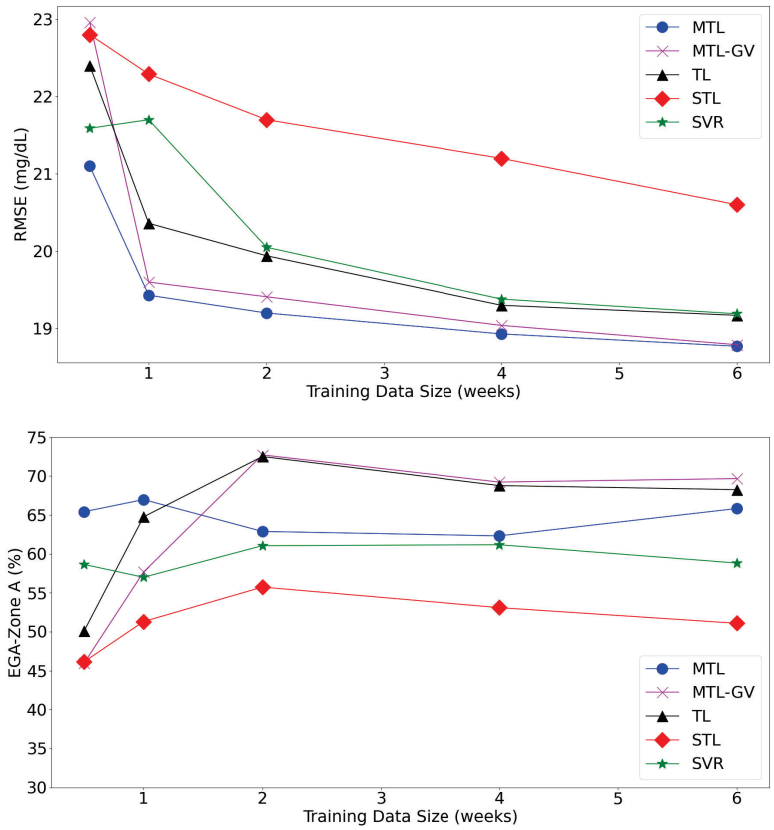

Figure 4.5: (Top) A comparison of the predictive accuracy, in terms of RMSE, for data-driven machine learning models and learning approaches at different training set sizes. (Bottom) The clinical relevance of errors in model prediction for hypoglycaemia at different training set sizes.

Table 4.3: Total number of samples at different training set sizes and associated samples in the hypoglycaemia region.

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

as the number training set size reduces. As seen in the first experiment, the clinical accuracy for all models in the safe regions is generally high i.e. Zone A-B $\geq$98%. As

seen in Table 4.3, the number of training points in the hypoglycaemia region (CGM $\leq$ 70mg/dL) are relatively scarce (3-4%), but very consequential for applications such as PLGS systems. The guidance for hypoglycaemia treatment is to ingest *rescue carbohydrates* and/or suspend the insulin basal rate to facilitate recovery of blood glucose concentration levels.

As seen in Figure 4.5a, multitask learning (MTL) provides the best performance in terms of RMSE at all training sizes. This is significant compared to the conventional TL approach (p-value $<$ .013). Furthermore, MTL are able to maintain this consistent performance in predictive accuracy when trained on at least 1 week of data (Δ RMSE $\leq$ 5%) from T1DM subjects.

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

Figure 4.5b shows the clinical relevance of errors in hypoglycaemia at each training set size and highlights the consistency in performance of the MTL approach. MTL approach shows better consistency (Δ_{max} Zone A = -4%) in performance over the various training sizes relative to the MTL-GV approach (Δ_{max} Zone A = -26.8%). In this scenario, MTL-GV models show the best performance when trained with at least 2 weeks of training data from each participant followed by the TL models. However, this performance drops sharply when training data size is further limited - Δ_{max} Zone A = -26.8% for MTL-GV and Δ_{max} Zone A = -22.4% for TL. The best performance in clinical accuracy for hypoglycaemia prediction is obtained by the MTL-GV model when the training data is reduced to the last 2 weeks of samples at 72.7%.

4.7 Discussion and future work

Multitask learning facilitates transfer of useful information between subjects. The results detailed in Section V demonstrate that the performance of the personalised

models improves with the introduction of population data, however, this benefit is more consistent with a multitask learning approach.

Small datasets can typically result in underperformance of deep learning methods which prompts the use of transfer learning[118]. One potential reason for the improvement in performance with the multitask learning approach, is that parameter sharing may serve as an additional form of regularisation which works to improve the generalisation of model performance on unseen data. This could also explain the high and consistent performance experienced, despite limiting the size of training data, relative to other approaches evaluated.

A clinical application that is highlighted in this work is the low-glucose suspend for artificial pancreas systems and hypoglycaemia alerts for clinical decision support systems. As noted by the results, the benefit of MTL is that it provides consistent performance with limited data in terms of hypoglycaemia detection. In addition, as the availability of training data increases over time ($\geq$2 weeks), clustering with respect to glycaemic variability (MTL-GV) and retraining models provides improved performance in hypoglycaemia detection. This results in the clinical benefit of deploying such tools earlier (relative to when trained with STL) without compromising performance.

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

Assessing the credibility of such models to be deployed in the healthcare domain is important and gaining attention. This credibility can be ascertained through model risk assessment, verification, and validation [142]. The validation and verification of the model can be considered through empirical metrics such as RMSE and MAE. On the other hand, model risk assessment can be evaluated on the EGA and temporal gain. For example, if consistent performance is sought by the clinician, the multitask learning (MTL) approach shows the most consistency in model performance even with reduced training data size. These considerations can give confidence to the clinician to recommend these models in a PLGS system even with limited subject data available.

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

4.8 Conclusion

Deep learning approaches are increasingly becoming relevant in developing the next-generation of diabetes management tools to aid in diabetes management. Glucose prediction represents a core part of that path, and as a result, effective methodologies are necessary to realise this with data-driven models. Deployment of such personalised models are hampered by the limited size of individual data available. Multitask learning provides an effective approach for leveraging population data to develop personalised glucose prediction models and overcome the challenge of scarce data for training models. Furthermore, incorporating prior information such

as glycaemic variability can be beneficial for long-term prediction tools in a multi-task setting. Finally, this approach is agnostic to the neural network architecture and can be compatible with other architectures developed in the future. The results from this work suggest multitask learning can facilitate a path for potentially deploying personalised models towards improving glycaemic control on limited individual data.

Chapter 5

Multitask Learning for Automatic Meal Detection and Estimation

Reproduced with permission from J. Daniels, P. Herrero, and P. Georgiou, "A Deep Learning Framework for Automatic Meal Detection and Estimation in Artificial Pancreas Systems" Sensors; published by Multidisciplinary Digital Publishing Institute, January 2022.

5.1 Introduction

The self-management of diabetes is a burdensome, delicate, and yet critical undertaking across different facets that individuals with diabetes have to engage in daily in order to avert adverse glycaemic events. Current diabetes management systems such as the artificial pancreas (AP) and decision support systems have been developed in recent years in order to improve the management of diabetes. The artificial pancreas is a system that comprises a continuous glucose monitor (CGM), insulin pump, and an algorithm working in tandem to maintain blood glucose levels in an acceptable range (70-180 mg/dL). The envisaged endpoint in the development of the insulin-based artificial pancreas is a fully automated system that does is not depend on user input throughout the day [36]. One major challenge in the realising this objective with the artificial pancreas centres on postprandial glucose control.

Studies have shown that when small meals (eg. 30g) are missed AP systems are capable of handling the resulting postprandial increase in glucose [143, 144]. However, control is noticeably poorer with larger sized meals [143, 144]. As a result, the current artificial pancreas systems is classified as a hybrid closed-loop system since it requires meals to be announced prior to mealtime in order to ensure good control.

The need for user attention to initiate meal announcements in hybrid closed-loop systems can lead to sub-optimal outcomes. In the ideal scenario, the user accurately estimates the meal size by carbohydrate counting and provides that information to the AP system prior to eating. This has to be done due to the delays associated with subcutaneous insulin delivery. However, multiple studies have shown that individuals with diabetes tend to have significant rates of late or missed meal insulin boluses [145]. This can be attributed to factors such as diabetes distress, stress and forgetfulness among others in a daily routine [146, 147, 148]. These studies generally showed a strong link between late and missed meal boluses and HbA1c levels. This increase in HbA1c levels can lead to worsening quality of life over time [149, 150, 46, 151, 152]. Furthermore, individuals with diabetes are reported to misestimate meal sizes between 20% and 59% [153, 154].

This appropriately motivates the need for systems such as the AP to be able to detect unannounced meals and estimate the carbohydrate size in order to control postprandial hyperglycaemia.

5.2 Related Work

A number of methods have been proposed in the literature towards tackling the detection and estimation of unannounced meals. These methods broadly fall into the following categories: (i) threshold-based detection using the rate of change (ROC) of glucose levels and (ii) outlier detection using model predictions.

Various threshold-based approaches are proposed that analyse a number of rules regarding the rate of change for meal detection. Dassau *et al.* [71] propose an

ensemble approach that comprises three or four approaches - the backward difference (BD), the second derivative of glucose, a Kalman filter (KF) estimation, and a combination of BD and KF - with meal detection dependent on attaining a majority vote from the approaches. Ramkisson *et al.* [49] employ an unscented Kalman filter that extends Bergmans minimal model to account for a disturbance parameter and tracks a cross-covariance of the forward difference of this parameter and glucose level, and the rate of change to confirm meal detection. To include meal size estimation, Samadi *et al.* [60, 47] initially analyse the first and second derivatives of a filtered CGM signal in order to detect a meal, and then estimates the meal size from glucose levels and insulin-on-board (IOB) using a fuzzy logic system.

A potential downside of relying on the rate of change and pre-defined thresholds for meal detection is that when there is significant variability in glucose levels this would lead to a reduction in the signal-to-noise ratio, and consequently, a relatively high rate of false positives which can be potentially harmful in this context [60, 47]. As a result, some approaches typically require filtering which may then increase the detection delay and reduce sensitivity. Even then, the thresholds are selected based on a training dataset with specific underlying conditions (eg. sensor noise, sensor drift, insulin sensitivity), so when these conditions shift significantly over the deployment period this may lead to degraded performance while going undetected.

The alternative approach is model-based approaches that then detect meals based on outliers in the glucose trajectory for which various approaches have been proposed as well [155, 48, 50]. Cameron *et al.* [155] develop a probabilistic approach that compares the expected signal and observed signal to detect a meal, and with assumptions on the meal shapes, estimate the glucose rate of appearance. A variable state dimension (VSD) approach is introduced where an extended Kalman filter (EKF) is used to predict the glucose trajectory with a 95% prediction interval [48]. A meal is detected, once the upper bound is exceeded, and a least-squares approach estimates the meal size. Mahmoudi *et al.* propose a similar approach that instead uses an unscented Kalman filter (UKF) to predict the glucose trajectory [50]. How-

ever this work employs two CGM sensors as opposed to one CGM sensor, and a meal is detected once both glucose trajectory from both sensors exceed the 95% prediction interval. Zheng *et al.* [156] employ the minimal model and run multiple simulations when the difference between the model trajectory and CGM trajectory is larger than a set value, to explain the likely meal size to reduce this divergence. Finally, Garcia-Torado *et al.* [73] rely on a classification model (logistic regression) that continuously estimates the probability of an unannounced meal and provides a scaled insulin bolus based on the total daily insulin instead of estimating the carbohydrate content.

This work devises a data-driven approach that leverages a deep multitask learning framework in order to detect and estimate meals. To the best of our knowledge this is the first meal detection and estimation framework based on multitask quantile regression, where the assumption on distribution of errors is relaxed. As a result we make the following contributions:

- We develop a novel meal detection algorithm based on multitask neural networks and quantile regression in order to automatically announce meals and estimate meal size.
- We evaluate *in-silico* the performance of the meal detection and estimation algorithm in moving towards fully closed-loop insulin delivery.

5.3 Methods

In this section we first present a meal detection and estimation algorithm that is based on a sequence to sequence model that is extended to a multitask setting in order to perform multiple quantile regression. We first describe the recurrent neural network that makes up this framework. This method is evaluated using a *in-silico* dataset of 10 adult subjects generated through the UVa-Padova simulator. We then illustrate the use case in the AP setting and present results on *in-silico* validation.

5.3.1 Multitask Deep Neural Network

As mentioned earlier, the deep neural network is based on a multitask sequence-to-sequence model. The sequence-to-sequence (seq2seq) model is primarily a model that is used to map one set of input sequences to an associated set of output sequences and has been used in glucose prediction tasks [157, 158, 126]. In this task, the objective is to estimate the last 20 minutes of the individual's glucose trajectory using historical CGM measures, meals, and insulin.

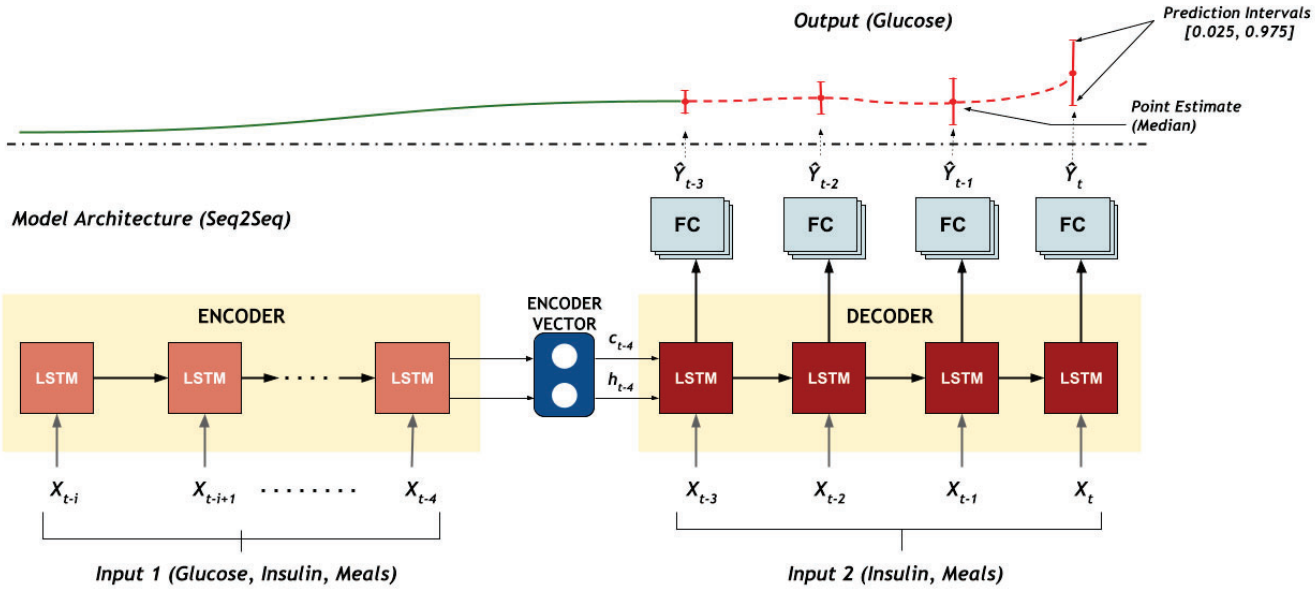

Figure 5.1: The multitask deep neural network architecture for predicting the 20 minute glucose trajectory at multiple quantiles. The multitask seq2seq model predicts glucose trajectory with a 95% prediction interval (PI) with a lower bound (2.5%) and upper bound (97.5%) .

To fully utilise the information available, a recurrent neural network encoder-decoder architecture shown in Figure 5.1 is employed in this framework to estimate the glucose trajectory. In this model, the encoder and decoder are based on long short-term memory (LSTM) networks due to their ability to better model sequential data without the issue of vanishing gradients [159]. The encoder LSTM input sequence - $\mathbf{x}_{enc} = (x_{t-i}, ..., x_{t-4})$ - comprises the glucose concentration levels from the CGM, insulin delivered and meals and returns the encoder state representations $\{\mathbf{c}, \mathbf{h}\}$. The decoder generates the output sequence - $\mathbf{y}_{dec} = (y_{t-3}, ..., y_t)$ - estimated quantiles

of the glucose trajectory are generated at the output. The decoder input, $\mathbf{x}_{dec} = (x_{t-3}, ..., x_t)$, comprises the insulin delivered and announced/estimated meals and is initialised with the final encoder state.

The output sequence generated from the decoder is then fed to the three final layers of the model that consists of three separate tasks. Each task represents a quantile distribution, τ, and consequently the model performs a quantile regression for the associated quantile.

Quantile Regression

In the earlier work by shown in Chapter 3 and 4, the deep learning model is trained using a mean absolute loss; this loss minimises the sum of absolute differences between the predicted value and the target value to perform a regression. In this work, we extend the utility of to output the glucose trajectory multiple quantiles.

Aleatoric uncertainty captures the uncertainty in the available data for training a model. This uncertainty may arise due to errors or variability in the data. In this case the errors may arise from the misestimation of meals or noise in measurements of glucose concentration levels from the CGM. As a result, once a persistent This forms the basis of the proposed approach to the meal detection and estimation framework in this work.

Quantile regression can be defined as method of estimating a conditional quantile, τ, where $\tau \in \lceil 0, 1 \rceil$ for a target variable, $\mathcal{Y}$ conditioned on the input, $\mathcal{X}$. Subsequently, the resulting prediction interval serves as an approach for estimating the aleatoric uncertainty in the training data. One approach that we utilise in training a deep neural network to perform quantile regression is the pinball loss/tilted loss as shown in Equation 5.1.

$$\mathcal{L}_\tau(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}), & y - \hat{y} > 0, \\ (1 - \tau)(y - \hat{y}) & \text{otherwise}, \end{cases} \tag{5.1}$$

where y refers to the reference glucose value, $\hat{y}$ refers to the predicted value, and τ refers to the quantile that the regression is estimating.

In our multitask architecture, each output corresponds to the following quantiles: a lower bound quantile (τ_{LB}), a median quantile ($\tau_M = 0.5$), and an upper bound quantile (τ_{UB}). The total of the quantile losses, $\mathcal{L}_{total}$, are uniformly weighted and jointly minimised as seen in Equation 5.2. It should be noted that the median quantile (conditional mean) is equivalent to the mean absolute error.

$$\mathcal{L}_{total} = \frac{1}{3}\lceil \mathcal{L}_{LB} + \mathcal{L}_M + \mathcal{L}_{UB} \rceil \tag{5.2}$$

The multitask architecture is beneficial in terms of computation and memory since the majority of weights are shared as opposed to having independent models for each quantile. Furthermore, *Tagasovska et al.* show that using a multitask architecture also enables consistent uncertainty quantification when modelling aleatoric noise in data [160]. Figure 4.6.3 below shows that consistent prediction interval of the from our multitask model for an adult subject.

Network Training

Prior to training the input features are first normalised. The neural network for each individual is trained in a two step strategy. The training dataset is split into 80% training data and 20% validation data. In order to attain better performance we first train a generalised model using aggregated data from each individual. The individualised model for each participant is then obtained by fine-tuning the generalised model on the individual data. Further details on the hyperparameters and optimisation of the model are included in Appendix C.

5.3.2 Meal Detection and Estimation

Meal Detection

The first stage of the model involves the detection of a meal. First, the last 20 minutes of the glucose profile is estimated using the sequence to sequence model. As explained in the previous section, the model outputs multiple quantiles for a 95% prediction interval coverage that arises from the errors arising from the input variables. Consequently, in a scenario where the glucose trajectory persists outside the prediction interval, this implies a significant *error* beyond noise in the input variables that can be inferred as a missing input. In this setting, this missing input is attributed to an unannounced ingested meal and therefore a meal detection flag can be activated.

The use case of this framework necessitates the priority of safety. As a result, in the detection of the algorithm, we set the condition that this *error* needs to persist for m samples to satisfy activating the meal detection flag. In addition, the flag is only activated if the rate of increase of glucose trajectory is at least 1 mg/dL/min.

Carbohydrate Estimation

Once the meal detection flag is activated we begin the process of estimating the meal size. This step is achieved with a simple iterative search approach implemented at the input in order to determine the best possible meal size to estimate the present CGM value. We condition this iterative search on the mean absolute error (MAE) between the median quantile of the present estimated glucose level and the reference glucose level from the CGM. The search is completed once the mean absolute error is less than a threshold, ϵ. We initially increase the meal input at 10g and review the mean absolute error (MAE) until the result is lower than the threshold. We then reduce the meal increment to 1g and rerun the last iteration to obtain a more precise meal size estimate. The maximum meal estimate, m_{MAX}, is set to 90g at a time to mitigate risk of overestimation; additional successive meal estimates can be

made to supplement the initial estimate.

$$\begin{aligned} \epsilon &= MAE_{val} \\ &= \frac{1}{N}\sum_{k=1}^{N} |y(k) - \hat{y}_M(k)|, \end{aligned} \tag{5.3}$$

where $\hat{y}_M(k)$ denotes the predicted glucose level for a given sample at the media quantile, k. $y(k)$ denotes the reference glucose measurement, N refers to the number of samples in the validation set.

For robustness and verification, a check is included where if the meal size increment does not lead to the expected physiological increase in trajectory, the meal is discarded. The algorithm pseudocode is detailed in Alg. 1. On a 24 hour glucose profile, a graphical representation of Alg.1 of the detection and estimation of meals is shown in Figure 5.2

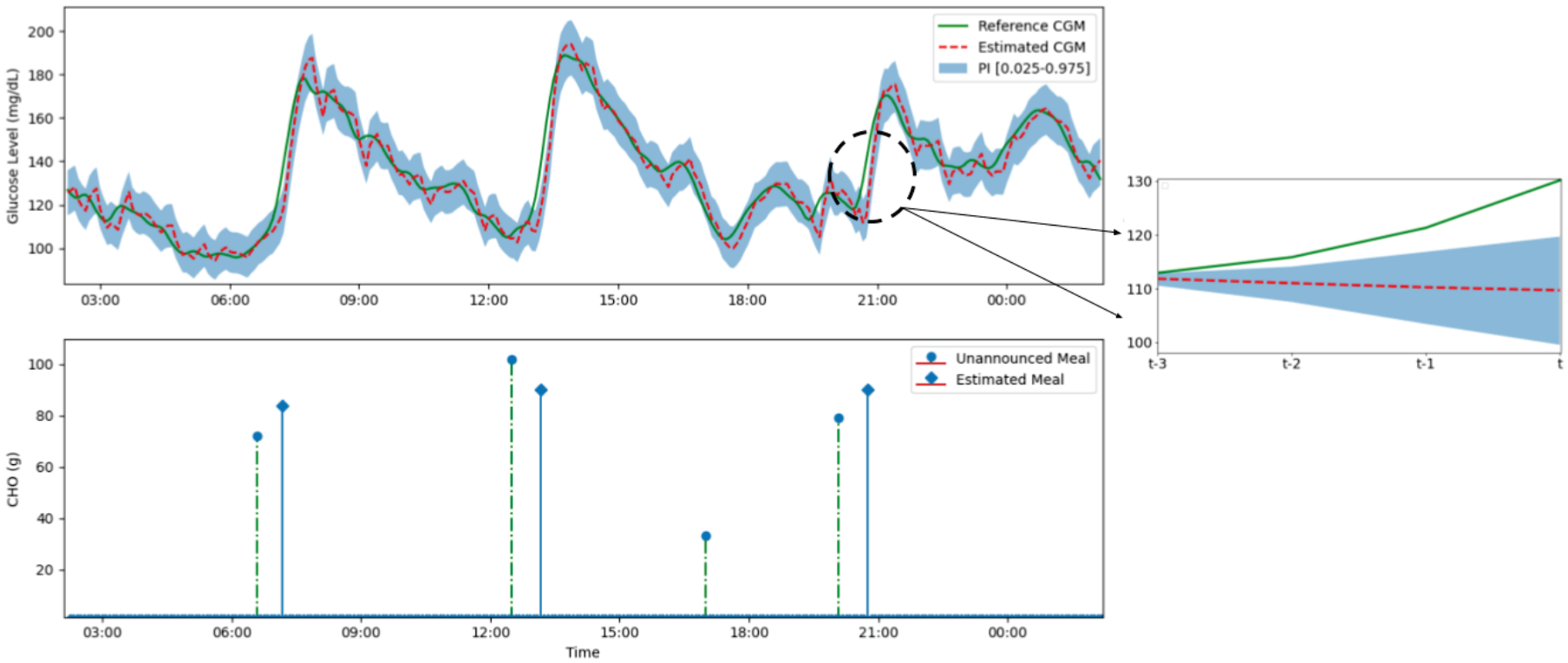

Figure 5.2: (Top) A figure showing the multitask seq2seq model predicts the 20 minute glucose trajectory for a 95% prediction interval (lightly shaded blue area) over a 24-hour period. (Bottom) The unannounced meal and reconstructed estimated meal. Shortly after unannounced meal a significant and persistent deviation in glucose trajectory from estimated trajectory leads to a meal detection and estimation. Certain meals such as the small snack between 3pm and 6pm are not detected as reliably since they only result in a non-significant deviation in the glucose trajectory.

Algorithm 1 Meal Detection and Estimation Algorithm

Require: $\mathbf{X} = \{\mathbf{G}, \mathbf{I}, \mathbf{M}\}$, ϵ
Ensure: Output meal size, m_E
  Initialise model parameters θ from memory
  $N \leftarrow 2$
  for $t \in 28, ..., T$ **do**
    $[\mathbf{y}_u, \mathbf{y}_m, \mathbf{y}_l] \leftarrow f(\mathbf{X}|\theta)$
    Compare $\mathbf{y}_u$ and $\mathbf{y}_{CGM}$
    Let k be number of samples where $\mathbf{y}_u < \mathbf{y}_{CGM}$
    if k > N `AND` ΔBG/Δt ≥ 1 **then** ▷ Activate meal detection
      $m_E \leftarrow 0$
      $error \leftarrow \texttt{Abs}(\mathbf{y}_m(t) - \mathbf{y}_{CGM}(t))$
      while error $> \epsilon$ `AND` $m_E < m_{MAX}$ **do** ▷ Perform meal estimation
        if `fine_search` **then**
          $m_E \leftarrow m_E + 1$
        else
          $m_E \leftarrow m_E + 10$
        end if
        $\mathbf{M} \leftarrow m_E$
        $[\mathbf{y}_u, \mathbf{y}_m, \mathbf{y}_l] \leftarrow f(\mathbf{X}|\theta)$
        $error \leftarrow \texttt{Abs}(\mathbf{y}_m(t) - \mathbf{y}_{CGM}(t))$
        if error $\leq \epsilon$ **then**
          $m_E \leftarrow m_E - 10$
          Activate `fine_search`
        end if
        if verify appropriate glucose dynamics **then**
          $m_E \leftarrow 0$ ▷ Discard meal estimate
          break
        end if
      end while
    else
      $m_E \leftarrow 0$
    end if
  end for

5.3.3 Fully Closed-Loop Control for Insulin Delivery

As mentioned in Chapter 2, a number of closed-loop controllers have been proposed in the literature in order to facilitate tight glycaemic control. Examples of the different types include proportional-integrative-derivative (PID), model predictive control (MPC), reinforcement learning-based (RL) controllers, and bio-inspired controllers. The bio-inspired artificial pancreas (BiAP) is a hybrid glucose controller based on

that has been extensively validated in previous work [144] and is implemented in this work to evaluate the algorithm. It should be noted, however, that this algorithm is agnostic to the choice of controller.

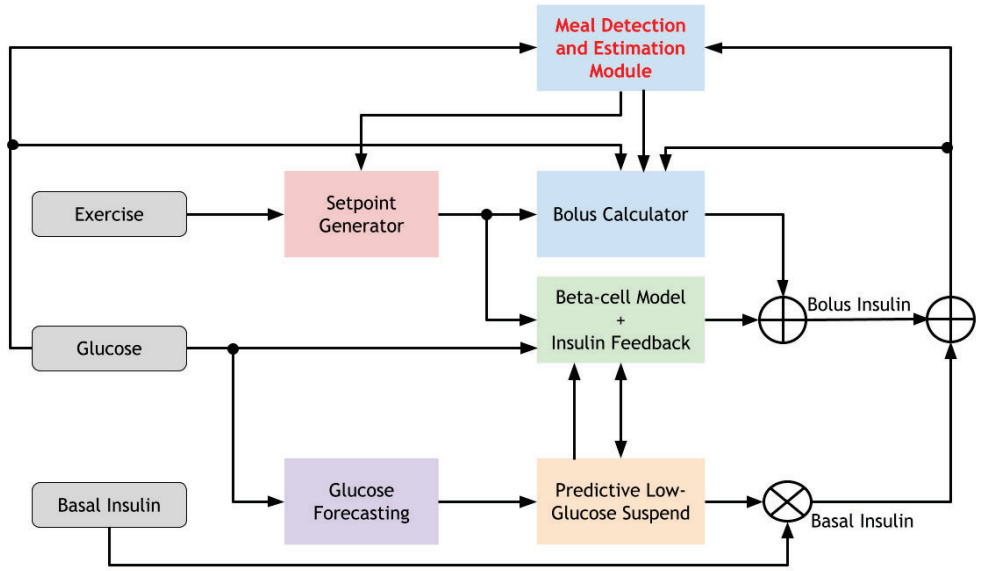

Figure 5.3: The system architecture of the bio-inspired artificial pancreas with the meal detection and estimation algorithm incorporated.

As shown in Figure 5.3 the meal detection and estimation module provides an estimate of the carbohydrate size to the bolus calculator in the BiAP controller to determine the meal insulin bolus. However, based on the initial increasing trend in the glucose trajectory the controller already begins to deliver the insulin boluses due to this deviation from the target glucose concentration level. This insulin is delivered as the deviation remains below the upper bound and the meal flag is not raised. The prudent measure in this case is to remain conservative and minimise the possibility of precipitating a postprandial hypoglycaemic event. Consequently a weight ($\mathcal{W}$=0.5) is applied to the calculated meal bolus before a final bolus is delivered, and the basal insulin delivery is suspended after for 1 hour.

5.3.4 Performance Metrics

We employ multiple metrics to comprehensively evaluate the meal detection and estimation framework. To assess the detection of meals we use the following metrics: precision, recall, false positive rate, and median delay.

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.5}$$

A true positive (TP) is identified when the detection flag is raised and the delay in identifying the meal is less than 120 minutes. Otherwise, this event is identified as a false negative (FN). A false positive (FP) is identified when a meal flag is raised in the absence of a meal. In the eventual use case of a meal detection algorithm, the detection of an unannounced meal would prompt a bolus to be delivered either indirectly by notifying the user with an alert, or directly in a sensor augmented pump. Subsequently, it is important that the metrics not only assess the effectiveness of the algorithm in detecting unannounced meals, but also we assess the quality of the meal detection. The recall (sensitivity) assesses the ability of the algorithm is detect a meal, whereas the precision assesses the quality of meal detection. The specificity is not considered in evaluating the framework since TN $\gg$ FP, leading to the algorithm posting extremely high values for specificity. For the meal detection time, we evaluate based on the median delay between the detected meals and the actual meals in the testing dataset. For assessing the effectiveness of *CHO* estimation, we indicate the mean error size and percentage error.

The performance of glycaemic control is evaluated using a comprehensive set of metrics that are typically used in the literature. We primarily report the following glycaemic metrics: percentage time spent in euglycaemia (70mg/dL <BG< 180mg/dL), percentage time spent in hyperglycaemia (BG$\geq$180mg/dL), percentage

time spent in hypoglycaemia (BG≤70mg/dL), mean glucose concentration level. We evaluate the level of control with a number of indices, particularly, the high blood glucose index (HBGI), low blood glucose index (LBGI), and risk index (RI). In addition, we consider the insulin per day used to achieve the associated level of control.

Finally, we provide a visual comparison of the quality of closed-loop glycaemic control of the different configurations with a control-variability grid analysis (CVGA) [161]. This visualisation is complemented with a numeric assessment of the quality of control.

Statistical Analysis

We evaluate the differences in the controller performance with different configurations: meal announcement, meal detection, and unannounced. For determining the statistical significance, we first perform a preliminary test for normality using a Shapiro-Wilk test. We use a paired t-test if normality is accepted, and a Wilcoxon signed-rank test when normality is rejected. Significance level is set at p-value $<$ .05. For multiple pairwise comparisons, we adjust the significance level to p-value $<$ .025 using Bonferroni correction. The data from the results are presented as Mean $\pm$ SD.

5.3.5 *In-Silico* Dataset

The University of Virginia/Padova (UVa/Padova) T1D Simulator is used to generate a challenging scenario for training, validating and testing the models. For the meal protocol scenario we choose four meals with the following average carbohydrate size at the associated average meal times: 40g (7am), 70g (1pm), 30g (5pm), and 100g (8pm). A meal-time variability (σ_T = 60 min) and meal size variability (CV = 10%) is introduced in order to generate realistic scenario of inter-day variability in meals. In addition, to account for additional variability in meal composition, the

simulator meal library was supplemented with a further 16 new meals as described by the authors in [162].

To generate more realistic scenarios, additional intra-day variability on insulin absorption and insulin sensitivity was introduced. Variability of insulin absorption is assumed to be ±30% and the insulin sensitivity varies in a sinusoidal manner with a selected daily period.

As mentioned earlier, studies have shown that individuals with diabetes are not always prompt with meal announcements and also the carbohydrate counting is consistently misestimated. In order to model this behaviour, a trigger for skipping and delaying meal announcement was randomly generated based on an average 2.5 meal announcements skipped, and 2 meals delayed per week. Carbohydrate counting uncertainty is incorporated in the simulator [-30%, +10%] with a uniform distribution. The values are selected due to the bias towards underestimation rather than overestimation of carbohydrate size. These parameter choices are then used consistently across all tests and comparisons.

5.4 Results

5.4.1 Meal Detection and Estimation

In this section we report the results on the performance of the framework in detecting and estimating the unannounced meals in the simulation scenario. The performance is reported on each mealtime and for the snacks.

Table 5.1: Cohort performance metrics for the meal detection performance for the different mealtimes. The meals section involves accounting for algorithm performance without snacks included.

Metric	CHO(g)				
	Breakfast	Lunch	Snack	Dinner	Overall (Meals)
	70±7	100±10	30±3	80±7	70±27 (83±15)
Meal Detection Performance					
Precision (%)	86±7	98±2	97± 5	94±6	93±4 (92±4)
Recall (%)	90±5	97±3	24±14	89±4	76±5 (92±3)
Delay (min)	38±13	36±11	41±23	37±15	38±15 (37±13)

Table 5.1 shows the meal detection performance of the meals and snacks at different mealtimes. The average breakfast meal size is of moderate size (70 ± 7g). The method obtains a precision 86±7% and a recall 90±5%. In terms of detection time these meals are flagged at 38±13 min. The lunch meals are the largest size considered in this study at 100±10g. At lunchtime, the proposed method detects lunch with 98±2% precision and 97±3% recall. This detection is completed in 37±15 min. Finally, for dinner, the average size during this mealtime is 80±7g which can also be considered moderate-sized. For this proposed method, we detect dinner meals at 94±5% precision and 89±4% recall. The snacks ingested after lunch are of a relatively smaller size at 30±3g. Although the precision is high at 97±5%, the recall is relatively low at 24±15%. In addition, the detection time is 41±23 min. The observation is made that the recall and detection time performance is dependent on the size of the meal, where small meals (snacks) show the worst performance and large meals (lunch) show the best performance, with moderately sized meals (breakfast and dinner) showing intermediate performance. On the other hand, we notice that the performance in terms of precision is consistently high (86-98%) across

different meal sizes.

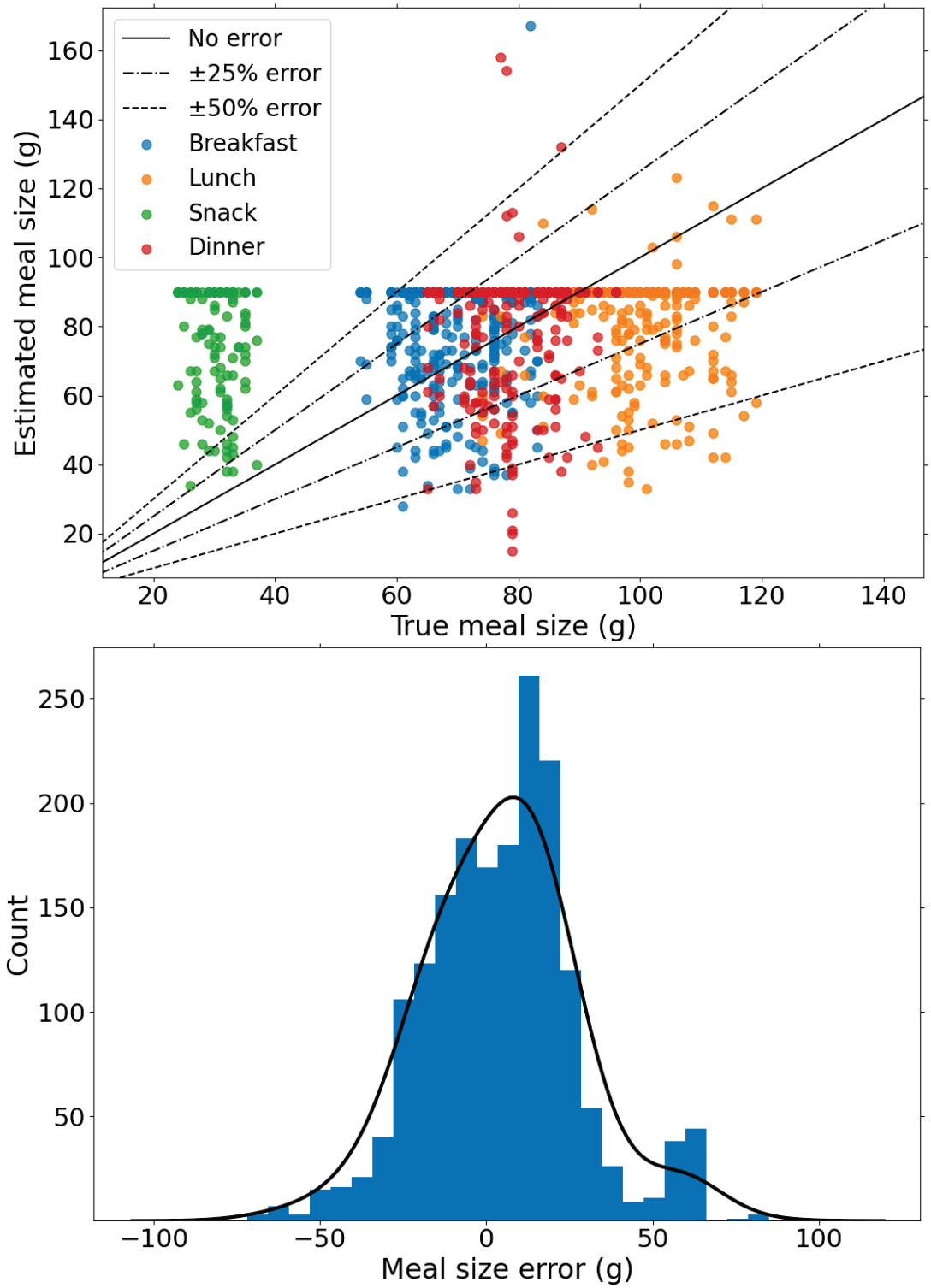

Figure 5.4: **Top:** Estimated meal size versus true meal size for detected meals to determine the accuracy of meal size error. **Bottom:** The distribution of estimated meal sizes and true meal sizes. The probability densities of meal and snack size errors is also shown.

Following the evaluation of the meal detection performance, we analyse the perfor-

mance of the proposed method on meal estimation. Figure 5.4 shows the comparison of the estimated meal size and the actual meal sizes and shows the distribution of estimation errors in meals. First, we note that meal estimation is estimated in relation to the median glucose level trajectory which have errors and thus can lead to under- and overestimation of meals. 80% of the estimated CHO have a estimation error within 25g, whereas only 6% of detected CHO is larger than 50g, of which snacks represent the majority. The distribution of the estimation errors, shown in Figure 5.4, shows that the proposed method is slightly biased towards overestimation with an average error 18±15g. This justifies applying the weight to the mealtime insulin bolus for the insulin delivery strategy to mitigate postprandial hyperglycaemia without significantly increasing hypoglycaemia risk.

5.4.2 Closed-loop Postprandial Glucose Control

In this section we report results of the performance of the BiAP controller with different configurations. The different configurations used in this study are described below:

BiAP-NMA: In this configuration, meals are not announced prior to the selected mealtimes for bolus priming. The controller is therefore only able to respond to the postprandial glucose excursion through feedback from the CGM signal. Since there is no external input from the user for meal announcement this is a fully closed-loop configuration.

BiAP-MD: This is a fully closed-loop configuration that corresponds to the BiAP controller with the meal detection and estimation module incorporated. In this configuration, the insulin bolus is delivered as explained in the closed-loop insulin delivery.

BiAP-MA: This hybrid closed-loop configuration corresponds to the controller with meal announcement included. Meal announcement involves the individual estimating the meal size and input this in the controller in order to deliver a preprandial

insulin bolus. The behaviour of the individual is modelled as earlier described to account for carbohydrate misestimation, missed boluses and late boluses.

Table 5.2: A comparison of the performance in terms of glycaemic metrics and risk indices between different configurations of the BiAP: meal announcement (BiAP-MA), meal detection (BiAP-MD), and without either detection or announcement (BiAP-NMA). p-values calculated with Wilcoxon signed-rank test are underlined.

Metric	Controller			p^*	$p^\dagger$
	BiAP-MA	BiAP-MD	BiAP-NMA		
Glycaemic Targets					
Mean BG (mg/dL)	137.7±5.0	144.5±6.8	148.9±9.8	<u>0.002</u>	0.003
TIR (%)	84.7±5.1	77.8±6.3	73.9±7.9	<u>0.002</u>	0.0007
TAR (%)	13.7±4.4	20.7±6.0	24.9±7.8	<u>0.002</u>	0.0009
TBR (%)	1.5±1.3	1.4±0.9	1.3±1.2	0.8	0.4
Risk Indices					
HBGI	3.2±0.8	4.3±1.1	5.1±1.5	<u>0.002</u>	0.0005
LBGI	0.5±0.4	0.6±0.4	0.5±0.3	0.3	0.1
RI	3.7±1.0	4.9±1.3	5.6±1.6	<u>0.002</u>	0.002

p^* = BiAP-MD vs. BiAP-MA; $p^\dagger$ = BiAP-MD vs. BiAP-NMA

The performance of the controllers in enabling tight glycaemic control is reported in Table 5.2. BiAP-MD and BiAP are the two closed-loop controllers that are described as fully closed-loop. A comparison of the performance between these two controllers reveals that the meal detection and estimation algorithm improves the control of postprandial hyperglycaemia. This is evident from the significant reduction in time spent in hyperglycaemia (ΔTAR = -4.2%, $p = 0.0009$) and lower risk of hyperglycaemia (ΔHBGI = -0.8%, $p = 0.0005$). Overall, BiAP-MD reports a significantly lower mean glucose level (-4.4 mg/dL, $p = 0.003$) than the BiAP controller and provides relatively tighter glycaemic control (ΔTIR = +3.9%, $p = 0.0007$). Finally, this is accomplished without a statistically significant increase in time spent in hypoglycaemia (ΔTBR = +0.1%, $p = 0.4$) or risk of hypoglycaemia (ΔLBGI = +0.1, $p = 0.1$).

A further comparison is made between the BiAP-MA and BiAP-MD controller. The first observation is that BiAP-MA has lower mean blood glucose level (-6.8 mg/dL, $p = 0.002$) than BiAP-MD. In addition, we see significant improvement in tight glycaemic control with BiAP-MA over the proposed BiAP-MD controller:

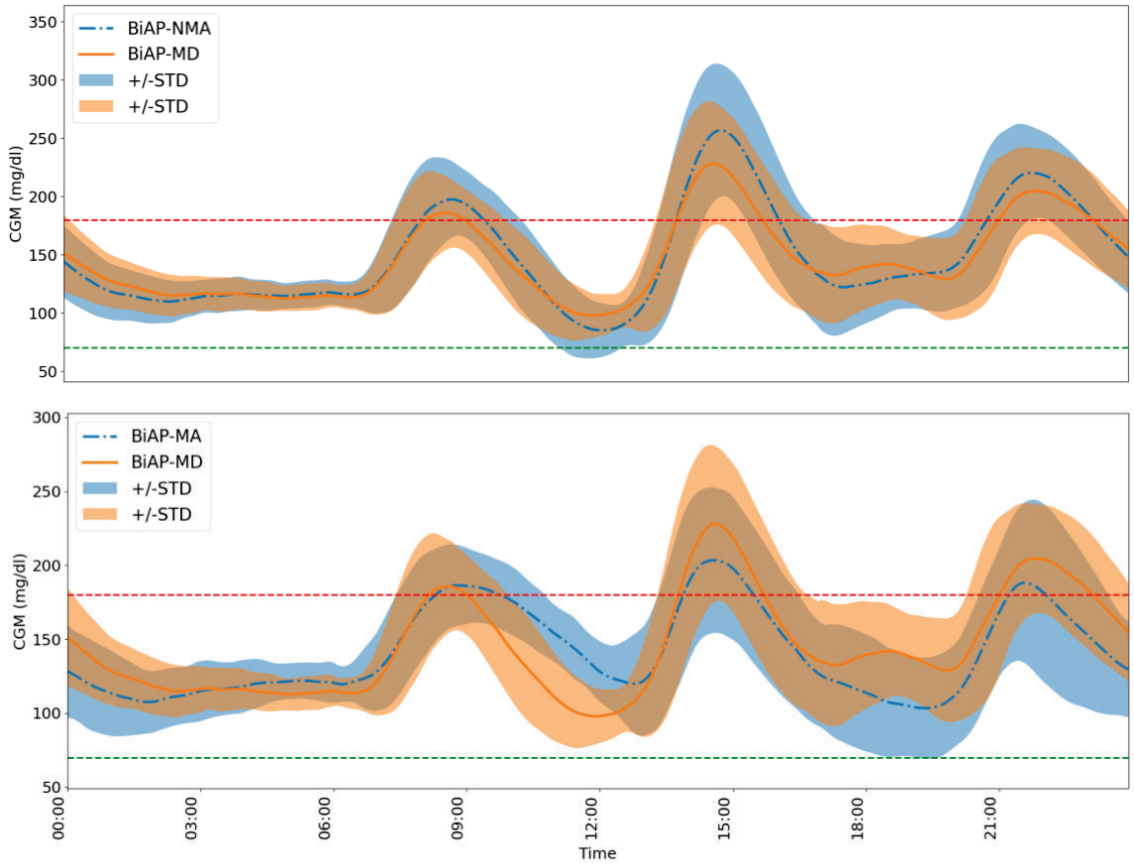

Figure 5.5: A comparison of the 24 hour glucose profile of Adult 3 from the virtual cohort over the 2 month period between the BiAP controller with meal detection and the two baseline configurations. The red and green lines mark the hyperglycaemia and hypoglycaemia threshold respectively. (Top) **BiAP vs. BiAP-MD:** A comparison between the BiAP controller performance without and with the meal detection and estimation incorporated. (Bottom) **BiAP-MA vs. BiAP-MD:** A comparison between the BiAP controller performance with user-initiated meal announcement and automatic meal detection and estimation.

increased time in range (+6.9% mg/dL, $p = 0.002$), reduced time spent in hyperglycaemia (-7%, $p = 0.002$), and reduced associated risk of hyperglycaemia (-1.1, $p = 0.002$). This difference in performance highlights the advantage of the individual pre-bolusing for meals over automatic meal detection and estimation. The accumulation of errors in meal announcements may lead to an increase in time spent in hypoglycaemia (+0.1%, $p = 0.8$) and associated risk of hypoglycaemia (+0.1%, $p = 0.8$), however, these are not observed to be statistically significant. The difference in 24 hour glucose profiles of an individual using the BiAP controller with the different configurations over the 2-month period is highlighted in Figure 5.4.2.

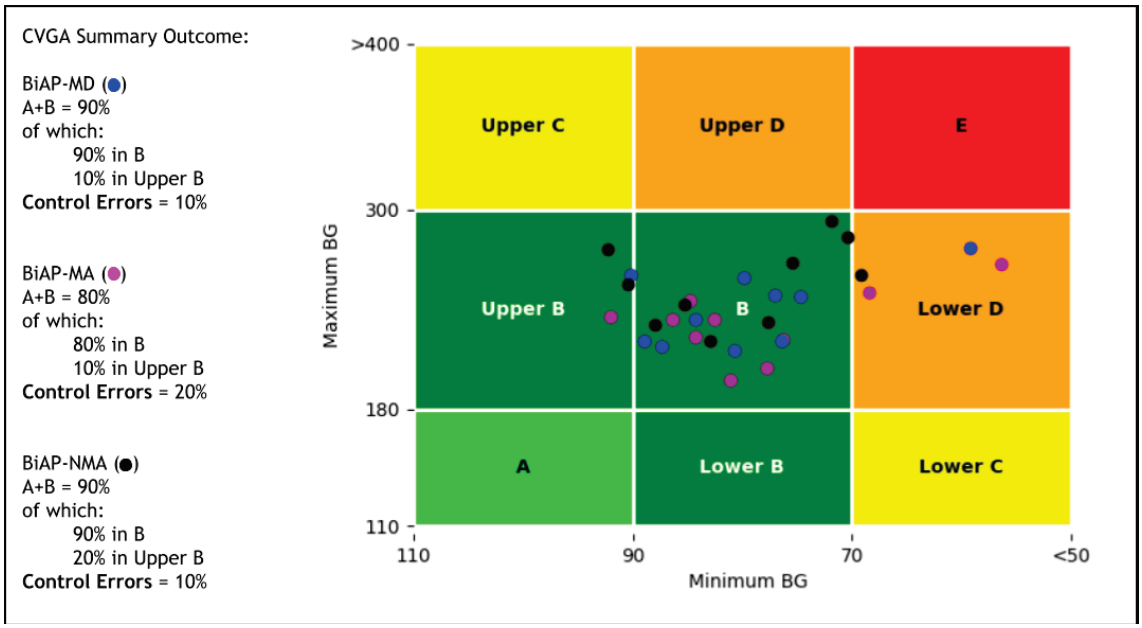

Figure 5.6: Control variability grid analysis for the BiAP controller with the different configurations reported. Each marker represents a virtual adult in the simulation assessed for the 2 month period.

An analysis of the CVGA plots on the population level shows a difference in the quality of glycaemic control for the different configurations. The first comparison we consider is between the fully closed-loop BiAP controllers (BiAP-NMA and BiAP-MD). The general numerical assessment shows that both configurations demonstrate a similar performance with 90% in Zone A+B and 10% in Lower D zone. In detail, however, the observation seen in Table 5.2 BiAP-MD exhibits tighter glycaemic control than BiAP-NMA is further buttressed in this plot. 10% of BiAP-MD markers were observed in the Upper B zone which is an improvement in comparison to 20% of BiAP-NMA markers, therefore displaying a lesser tendency towards benign control deviations into hyperglycaemia.

On the other hand, for the second comparison we consider quality of glycaemic control between BiAP-MA (hybrid closed-loop configuration) and BiAP-MD (fully closed-loop configuration). BiAP-MA shows marginally worse quality control with 80% of the population in Zone A+B compared to BiAP-MD with 90% in Zone A+B. As seen in Table 5.2 earlier, BiAP-MA provides tighter glycaemic control that BiAP-MD, however, this can is more likely to lead to more instances of hypoglycaemia

particularly when meals are overestimated and delayed. This would explain the higher instance of individuals from the population in Lower D zone and therefore a failure to deal with hypoglycaemia during control.

5.5 Discussion

5.5.1 Comparison with Other Approaches

As discussed in the related works section, a number of methods have been proposed towards detection and estimation of unannounced meals. These generally come under Kalman filters and/or heuristic rules. The reported metrics of the approaches are reported in Table 5.3 below.

Table 5.3: A comparison of reported performance metrics in the literature for automatic meal detection and estimation algorithms with our proposed approach.

Algorithm	Performance Metrics					
	Precision	Recall	F-Score	Delay	Size Error	UQ
Dassau et al.[71]	-	-	-	30 min	-	✗
Ramkissoon et al.[49]	92.5%	82%	0.87	38 min	-	✗
Samadi et al.[60]	79%	87%	0.86	-	23%	✗
Samadi et al.[47]	79%	93.5%	0.86	35 min	-	✗
Zheng et al.[156]	93%	88%	0.91	26 min	-	✗
Xie and Wang [48]	84%	76%	0.80	45 min	43%	✓
Mahmoudi et al.[50]	-	99.5%	-	58 min	-	✓
Ours - All	93%	76%	0.84	38 min	31%	✓
Ours - Meals	92%	92%	0.92	37 min	19%	✓

UQ = Uncertainty Quantification

We examine the performance of models that use heuristic rules such as inspecting the rate of change of glucose concentration levels. Dassau *et al.* [71] evaluated their meal detection algorithm on 17 subjects who consumed breakfast (22g - 105g). The CGM sampling time interval is 1 min. The detection time from meal onset the ensemble method is reported to be 30 min. This discrepancy in sampling time and reported metrics makes a fair comparison difficult.

Samadi *et al.* [60] studied an in-*silico* population of 30 individuals - comprising 10 adults, 10 adolescents, and 10 children. The overall performance is reported to be

91.7% precision and 91.3% recall. The meal detection time is not provided but the meal estimate error is 23.1%. However, the performance of the adult cohort is more comparable with the study undertaken in this chapter. Due to the higher variability in adults the results are lower, with 79% precision, 87% recall, and 22% meal size error. This is further supported by a study with 11 adult clinical subjects that showed 93.5% recall and 79% precision with a detection time delay of 34 min on average [47]. Zheng *et al.* [156] report on average a 88% recall and 93.3% precision with a detection delay time of 26 min when evaluated on 100 in-*silico* subjects. A meal size estimation error of 1.2±3.6g is reported, although this comparison is unfair as the meal sizes and size range evaluated on is relatively small (14-40.8g). Finally, Ramkissoon *et al.* [49] also evaluate their approach on 10 in-*silico* adults. Their trade-off setting, which is meant balance between false positives (FP) and recall, demonstrates 82% recall and 38 min detection time from meal start time. The authors report a false positive rate of 0.2 per day, and from the reported data a mean 92.5% precision is determined. However this approach does not estimate meals for enabling postprandial control.

The proposed approach is more comparable to those of Xie and Wang [48] and Mahmoudi *et al.*[50] as these rely on uncertainty quantification and outlier detection to detect unannounced meals. Mahmoudi *et al.*[50] study 10 adult subjects from an in-*silico* cohort. The reported recall (99.5%) for this method is relatively higher than our proposed method. However, this comes at the expense of the detection time as this is longer at 58 min, and the precision is not reported. Xie and Wang [48] evaluate the VSD performance with 30 in-*silico* participants. The reported recall (76%) and precision (84%) are relatively lower than the our proposed method. In addition, the meal detection delay is further (45 min) than our proposed method and has worse meal estimate error. Both approaches use a Kalman filter to quantify uncertainty although this assumes the distribution of errors is Gaussian and may be be the reason for the differences in performance. The primary difference in using multiple quantile regression as opposed to Kalman filters is that the assumption of

normality is relaxed.

Given the importance of the precision and recall, we compare the algorithms using the F-score which is the harmonic mean of the two metrics. Our proposed method achieves an average F-score of 0.84 for both meals and snacks. Given that meals are the primary challenge for automated postprandial glucose control, we also consider that when our approach is solely evaluated on meals - as is done with other methods - it achieves the highest average F-score of 0.92.

A limitation in comparing this work with other works across the literature is the difference in datasets used. The current clinical datasets available (OhioT1DM and ABC4D) were not suitable for testing as the meals are not always recorded, or at times recorded at a delayed time. Consequently, the stated metrics used in this chapter would not be accurate in evaluating the performance of this methodology. In future work, we intend to evaluate this work using both a suitable real-world clinical dataset.

5.5.2 Misestimation of Carbohydrate Content

One of the primary challenges in developing a suitable framework for detection and estimation of unannounced meals is the misestimation of carbohydrate content, either by underestimation or overestimation. As noted in Figure 5.2, the error in the model prediction varies i.e. the median glucose prediction level is sometimes higher or lower than the true CGM signal. Consequently, the use of the mean absolute error of the validation results in the possibility of underestimation or overestimation of carbohydrate content during meal estimation. In addition, different meals and snacks tend to lead to different glucose rates of appearance when ingested which may lead to misestimation if the model only learns a single glucose rate of appearance. The effect of this limitation is partly addressed by the 90g limit on estimated meals at any instant which explains the concentrated horizontal line observed in Figure 5.4.

In future work, the extent of misestimation can be minimised by improving the predictive performance of the model. This may be achieved by incorporating physiological factors – for example, explicitly allowing different rates of appearance of glucose from different meals/snacks – in our current approach. This would have the subsequent effect of reducing the extent of misestimation of carbohydrate sizes by reducing the error in predicted glucose levels.

5.5.3 Safety Monitoring with Uncertainty Quantification

Safety is an important factor in the development of automated insulin delivery systems. Consequently, safety considerations are generally considered not only in the evaluation but in the development and deployment of such systems as well. A prominent challenge once such data driven models are deployed is that a scenario such as dataset shift can lead to sub-optimal performance [163]. Dataset shift occurs when there is a change in the conditions present in the training setting. This can be the case when, for example, the behaviour of the individual and/or the CGM sensor noise is different in deployment. The result of this can be increase in the number of false negatives or in the worse case of increased variability as noted in [47], an increase in false positives. This could increase potential risk of hypoglycaemia.

Table 5.4: A comparison of glucose concentration levels for both false positive and false negative cases at different times during automated insulin delivery.

Failure	Blood Glucose Level (mg/dL)		
Case	At occurrence	+90 minutes	+180 minutes
False Positive	126 ± 33	120 ± 50	127 ± 55
False Negative	148 ± 38	151 ± 33	145 ± 36

In the proposed methodology, we observe from the precision and recall metrics that failure cases exist, although they are few. In general, the relatively higher precision compared to the recall corresponds to the presence of relatively more false negatives than false positives. For false negatives, this indicates undetected meals which mostly occurred with small carbohydrate sizes (<40g), and can be mostly attributed to the subsequent blood glucose rise being smaller than the estimated

size of prediction intervals. However, as noted earlier, the AP is capable of controlling postprandial glucose levels when carbohydrate size are smaller. This is further evident in Table 5.4 given that the average change in glucose level is relatively small over the 180 minute (3 hour) period from the missed meal.

For false positives, the primary observation is that the false positives tend to occur during high glucose levels. The false positives could be attributed to varying factors such as a decrease in insulin sensitivity which could then lead to a temporary glucose increase beyond the estimated upper bound of glucose concentration levels. As seen in Table 5.4, there is a decrease in the average glucose levels 90 minutes after a false positive detection. However, this change in blood glucose level is reduced after 180 minutes in this scenario given the intake of meals later on. This is further evidence of the safety in implementing the conservative approach for automated insulin delivery.

Although overcoming these moments of system failure remain an area of active research, our proposed method provides an avenue to monitor the possible cases of significant distribution shift. In deployment, the prediction intervals can be monitored over successive periods (eg.overnight) to detect if a significant deviation has occurred in the coverage of prediction intervals. This is possible given that the computed aleatoric uncertainty is generally unaffected by significant distribution shifts [164]. This behaviour can serve as a marker for when the deployed model needs to be retrained in the event that the model is not robust to such distribution shifts. As a result, uncertainty quantification should be an essential component for this application area moving forward and form the basis of future work for building safe and reliable systems.

5.6 Conclusions

Current artificial pancreas systems are hybrid closed-loop controllers and therefore require the user to perform manual meal announcements in order to adequately handle postprandial hyperglycaemia. Although preprandial bolusing is shown to be

most beneficial in achieving glycaemic targets, this places a cognitive burden on the user. Furthermore, the quality of glycaemic control is dependent on how well the individual is engaged with timely meal announcement and accurate carbohydrate estimation, which studies show is not always the case.

We develop a novel algorithm for meal detection and estimation of unannounced meals based on neural networks and multitask quantile regression. Compared to existing algorithms, this proposed approach achieves a better F-score for meal detection and competitive meal estimation performance based on simulation results. In addition, this algorithm provides a significant improvement in an artificial pancreas system to provide more effective closed-loop control. The hybrid closed-loop configuration shows better glycaemic control than our proposed approach, however, is worse at dealing with hypoglycaemia during control from the CVGA assessment. This study suggests that our proposed algorithm can serve as a viable approach for achieving fully automated closed-loop insulin delivery.

Chapter 6

Uncertainty-Aware Learning for Enhanced SMBG

6.1 Introduction

In Chapter 2, the proliferation of smartphones and the development of biosensors is discussed and noted to have led to devices and systems such as the continuous glucose monitor (CGM) and flash monitors that comprise the range of solutions geared towards enabling effective diabetes self-management. Studies have demonstrated the ability of these devices to improve tight glycaemic control, the primary approach for managing diabetes [165, 12].

Although these new devices show great promise in the effort to empower individuals in the self-management of diabetes, they are still not widespread across the world due to the cost and availability [7]. In addition, some individuals are unable to wear these devices for extended periods of time without experiencing pain and/or discomfort [166, 167, 168]. The myriad structural, economic, and physiological reasons that exist are the reasons that it is expected that the range of diabetes management tools outside of CGM-based solutions will persist for some time [7].

Glucose meters remain the most ubiquitous tool that enables individuals to monitor their blood glucose concentration levels. However, this only provides snapshots of

the glycaemic profile as opposed to CGM and flash monitors which also provide information on the trend, hence is not as effective for enabling glycaemic control [169, 170, 171]. In fact, it is supported in a number of studies performed that a greater SMBG sampling frequency is correlated with a reduction in HbA1c [172, 31]. Furthermore, it is suggested that real-time CGM which provides the best resolution and sampling frequency of glucose concentration levels is better than other intermittent glucose sensing alternatives [173].

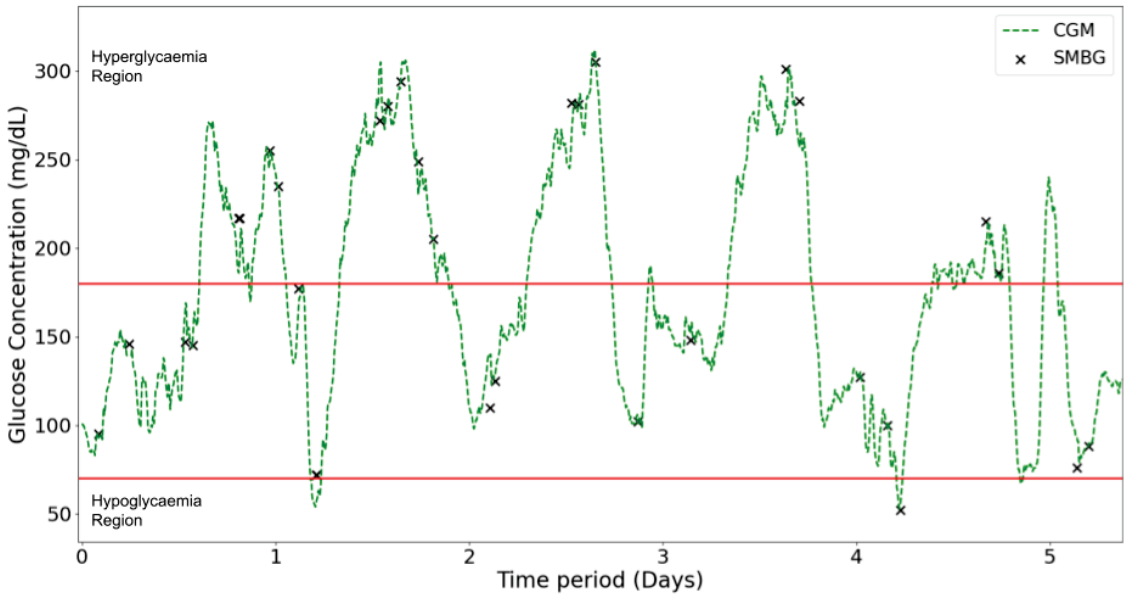

Figure 6.1: The variability of the glucose profile is best captured by the CGM as opposed to through SMBG [3].

Recent works in the literature have demonstrated that machine learning can be used to improve the capabilities of existing diabetes management tools [174, 33]. However, these works largely focus on the use of CGM which limits the wide applicability of machine learning for individuals with diabetes. As more applications based on machine learning methods are introduced towards improving the outcomes of one section of the diabetes population, this risks widening the chasm of healthcare inequality whereas machine learning can be employed in bridging this gap and facilitating equitable outcomes across the population.

The primary challenge in the extension of these data-driven approaches to devise tools for individuals with diabetes that use only standard care tools (glucometers

and insulin pens) is the sparse and irregular nature of fingerstick measures. The existing machine learning approaches used in these tools require structured regular samples as provided by CGMs - typically in 5 minute intervals.

This chapter introduces a machine learning approach that combines Gaussian processes (GP) and deep neural networks to develop an uncertainty-aware framework for monitoring glucose concentration levels in a more continuous manner. The main aim is to enable nowcasting of glucose concentration with trend information, even with sparse and irregular fingerstick measures provided with glucose meters. We address the following areas through our work:

- We study the performance of our uncertainty-aware learning approach against comparable methods and existing machine learning approaches.
- We investigate the improvement in detection of adverse glycaemic events over the current SMBG approach.

To the best of our knowledge, this is the first work that develops an uncertainty-aware machine learning framework for forecasting blood glucose concentration levels from glucometer readings rather than continuous glucose monitors in order to improve SMBG-based management.

6.2 Related Work

As noted earlier, few works exist in the literature that have leverage data-driven methods towards improving SMBG-based management. This can be attributed to the intermittent nature of this management routine and the subsequent emergence of real-time CGM devices which are more suited to current machine learning methods that are compatible with regular sampling intervals.

CADMO (Computer-Assisted Diabetes Monitor) is an early decision support system that is geared towards assisting healthcare professionals in providing care for insulin-dependent individuals with diabetes [40, 41]. The system mainly comprises

a mathematical model that simulates the glucose-insulin dynamics; the model estimates a continuous glucose profile by interpolating the glucose values obtained through the glucose meter, and in addition generates a continuous insulin profile based on the individual's MDI history. The combination of physiological, statistical, and rule-based methods then can be used to simulate the glucose trajectory and determine optimal insulin regime for good control. The limitations to this approach is the lack of personalisation as average values for parameters are used, hence some individuals that experience high variability in glucose and insulin therapy are not represented adequately. System recommendations are suggested to be less useful with more incomplete information. Similar to CADMO, KADIS (KArlsburg DIabetes management System) [39] is primarily centred around involvement of the clinician rather than primarily around the individual with diabetes.

On the other hand, another approach explored in the literature [175, 176] is the use of run-to-run control algorithms for optimising insulin delivery in order improve the time in range for T1D individuals. These proposed methods are geared towards individual use rather than clinicians for self-management.

Gu *et al.* [177] provide an alternative approach to complementing SMBG care with BGMonitor. BGMonitor is a decision support system that is developed using recurrent neural networks for classifying four glycaemic event classes. The four glycaemic event classes are Level 1 which corresponds to hypoglycaemia, Level 2 and Level 3 correspond to euglycemia, and Level 4 which corresponds to hyperglycaemia. The system then alerts the user to initiate a finger prick measurement and subsequently take a decision based on .

Joint models of Gaussian processes and deep learning models have been proposed in other areas as well in order to overcome the challenge of sparse and irregular inputs. The most popular application area is in sepsis management [178, 179, 180, 181], where this is used to process lab tests and vital signs to detect predict onset of sepsis which is primarily a classification task. Urteaga *et al.* [182] provides the closest method to our approach, however this approach is developed to model reproductive

hormonal dynamics in women. Another difference lies in the use of an attention-based neural network instead of a convolutional neural network.

6.3 Methods

6.3.1 Model Architecture

We detail the proposed approach, shown in Figure 6.2 below, which comprises two submodules: a Gaussian process (GP) for modelling the historical glucose concentration levels from sparse glucose meter measurements, and a deep neural network for forecasting the glucose prediction level from the resulting posterior distribution of glucose concentration levels, and the insulin, meals, and exercise logs. As seen in the previous chapters, recurrent neural networks are appropriate models for blood glucose prediction. This evidence is also supported in the literature with various recurrent network architecture designs being used for blood glucose prediction and beyond as well [117]. The deep neural network used in this framework is primarily an encoder-decoder architecture with self-attention, most similar with transformer architectures [183, 184].

Gaussian Process. A Gaussian process can be formally defined as a collection of random variables, any finite number of which have joint Gaussian distributions [185]. Given that modelling the historical glucose concentration (output) from the sparse fingerstick measures (input), this is framed as a regression problem where assumed a mapping exists between inputs and output. The Gaussian process can be described by the mean function $m(x)$ and covariance function, $k(x, x^{'})$ and subsequently, defines a probability distribution over possible functions, $f(X)$.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x^{'})) \tag{6.1}$$

For the purposes of simplicity, the the distribution is assumed to be zero-mean ($m(x)$=0). To construct the covariance function, we use a combination of kernel

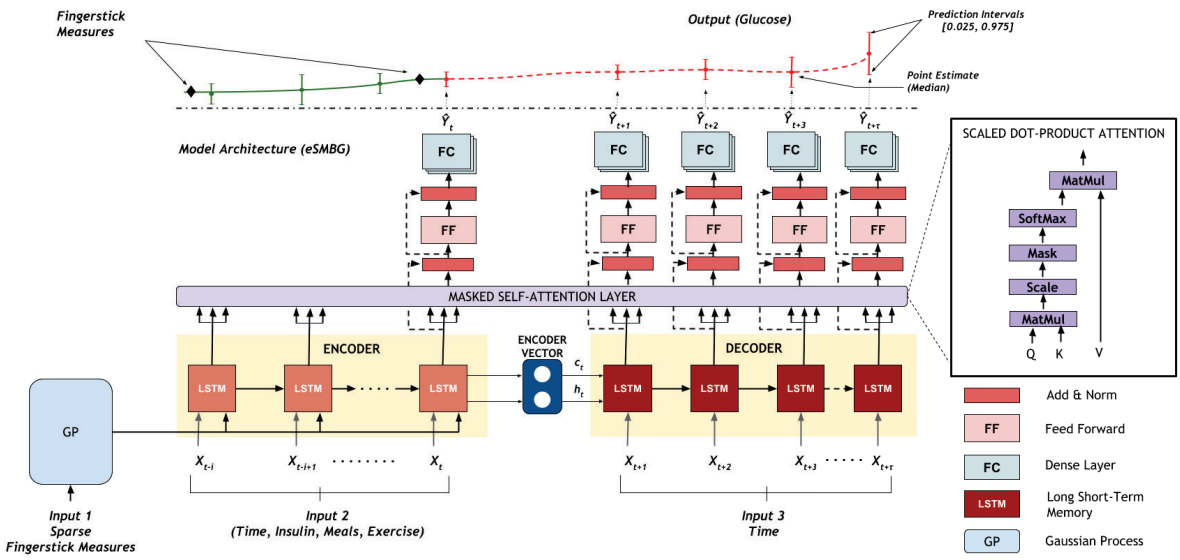

Figure 6.2: The proposed approach comprises a Gaussian Process for modelling the historical glucose concentration levels and a deep neural network for providing estimation and forecast of glucose prediction levels up to τ timesteps ahead to determine the trend.

functions in order to introduce appropriate prior knowledge about the glucose dynamics seen in the CGM profile.

$$
\begin{aligned}
k^*(x, x^{'}) &= k_A + k_B \\
&= k_{RQ}(x, x^{'}) \cdot k_{Per}(x, x^{'}) + k_{RBF}(x, x^{'}) \\
&= \sigma_1^2 \exp(-\frac{2 \cdot \sin^2(\frac{\pi}{p}|x - x^{'}|)}{\ell_1^2})(1 + \frac{(x - x^{'})^2}{2\alpha\ell_2^2})^{-\alpha} \\
&\quad + \sigma_2^2 \exp(-\frac{(x - x^{'})^2}{\ell_3^2}) + \sigma_n^2 \mathbf{I}
\end{aligned}
\tag{6.2}
$$

The covariance function as shown in Equation 6.2 is composed of two main components. k_A is a locally periodic kernel that is derived from the multiplication of a rational quadratic kernel(k_{RQ}) and a periodic kernel(k_{Per}). We initialise the period of the periodic kernel, $p = 1$ in order to model the long- and medium-term signal trends. .The smaller short-term trends are captured by the RBF kernel. Finally, the final component ($\sigma_n^2 \mathbf{I}$) is the white noise kernel that contributes the random noise associated with the glucose meter where σ_n^2 represents the variance noise and $\mathbb{I}$ is an

identity matrix. $\theta = \{\ell_1, \ell_2, \ell_3, \sigma_1, \sigma_2, \sigma_n, \alpha\}$ is the full set of hyperparameters that are learned during the training process.

The training process for learning the hyperparameters involves maximizing the marginal likelihood of the GP, $p(\mathbf{y} \mid X, \theta)$, from the training data (X, $\mathbf{y}$) as shown in Equation 6.3 below. Consequently, this corresponds to minimizing the negative log marginal likelihood to obtain the optimal hyperparameters.

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}}(p(\mathbf{y} \mid X, \theta)) \\ &= \underset{\theta}{\operatorname{argmin}} - \log p(\mathbf{y} \mid X, \theta) \end{aligned} \tag{6.3}$$

The output of the GP is a posterior distribution that enables interpolation of the sparse, irregular glucose concentration levels from fingerstick measures to samples with regular 5 minute intervals as seen in Figure 6.3.

This probabilistic approach provides a mean value ($\bar{f}$),

$$\bar{\mathbf{f}} = K_*(K + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{6.4}$$

along with a variance measure (Var(f)).

$$\mathbf{Var}[\mathbf{f}] = K_{**} - K_*^T(K + \sigma_n^2\mathbf{I})^{-1}K_* \tag{6.5}$$

where K refers to the covariance of the all training sample points, K_* refers to the covariance between the training sample points and the interpolated points, and K_{**} refers to the covariance between all the interpolated points in the observed space. Alternatively, the posterior can be sampled from the full covariance function to obtain multiple samples rather than use just the mean value as seen in Equation 6.6.

$$\begin{aligned}\mathbf{f} &= \bar{\mathbf{f}} + (\mathbf{Var}[\mathbf{f}])\mathbf{z}, \text{ where } \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &= K_*(K + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} + \text{chol}(K_{**} - K_*(K + \sigma_n^2\mathbf{I})^{-1}K_*)\mathbf{z}\end{aligned} \tag{6.6}$$

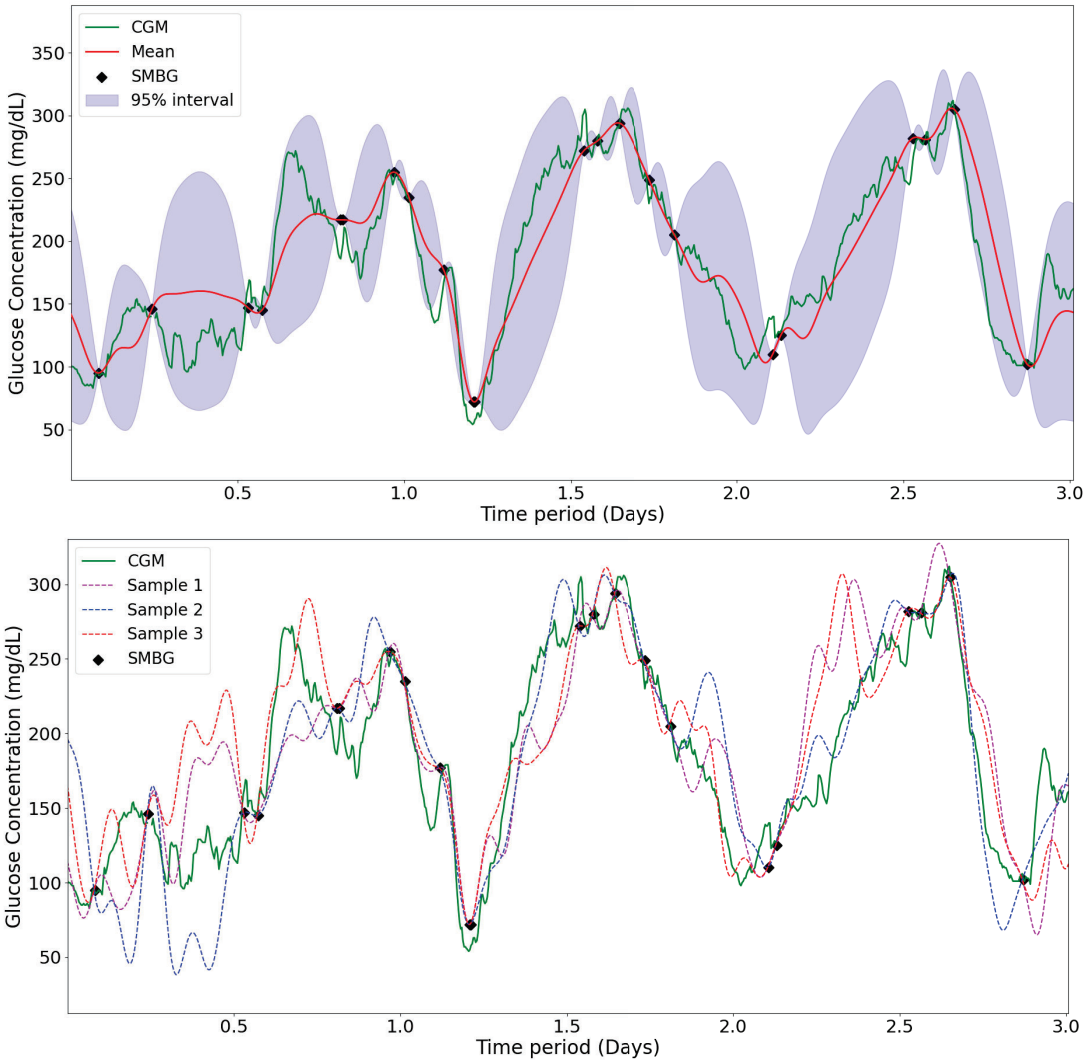

Figure 6.3: The Gaussian process regression is generated from the observed values from the observations collected from the fingerstick measures (SMBG) and compared to the CGM signal. **Top:** The mean of the posterior distribution and prediction intervals for the observed period. **Bottom:** Three posterior samples drawn and displayed from the posterior distribution for the glucose concentration levels.

Figure 6.3 shows the output of the GP. As we see, the uncertainty of the output increases the further away the interpolated value is from the observed glucose meter readings. The sparse glucose concentration levels are interpolated to obtain either the expected mean of the GP output or multiple posterior samples are generated at

regular intervals to then feed into the deep neural network.

LSTM Encoder-Decoder Layer. The initial layer of the deep neural network consists of an LSTM encoder-decoder architecture. The encoder processes the historical input of the reported data (exercise and meals), insulin, time of day, and the output from the GP of the glucose concentration levels over the last 2 hours. The final state and context vectors, h_t and c_t respectively, of the encoder feeds the decoder which only takes the time of day as an input (the only known future input at both training and test time). The output of the state, h_i, at each cell in the encoder and decoder are further processed by a residual layer and then fed into the self-attention layer.

Self-Attention Layer. The output of the sequence-to-sequence layer feeds into a self-attention layer that helps to learn long-term relationships from different points in the temporal sequence. The attention mechanism used in this work is the scaled dot-product attention, which takes the inputs as vectors: a query (Q), key (K), and value (V). This is shown in Equation below.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \cdot V, \tag{6.7}$$

where Q and K are vectors of dimension d_k, and V is a vector of dimension d_v. The output can be viewed as a scaled summation of the values based on the similarity of the query and key vectors. Although the transformer is typically implemented with multiple heads, the validation performance from the hyperparameter search resulted in implementation with a single head (context vector), $\mathbf{H}$. The self-attention layer is also masked in order to enforce a causal representation.

Position-wise Feed-forward Layer. The penultimate section of the model is the position-wise feed-forward layer that applies a feed-forward network to each position of the that further processes the output from the self-attention layer.

$$z = \max(e^x - 1, W_1 x + b_1) W_2 + b_2, \tag{6.8}$$

where W, b represent the weights and biases of the dense layers in the position-wise feed-forward layer. z represents the output of this layer after processing by an ELU activation function. This architecture also features skip connections between the LSTM layer and position-wise feed-forward layer to dynamically minimise model complexity where necessary.

Quantile Output. The final layer of the deep neural network is a multi-output layer in order to setup a multitask learning framework where each task corresponds to a pre-specified quantile, q. For each quantile, the prediction is made from t=0, to t=τ which is the furthest forecast value. This serves as a way to communicate the uncertainty from the inputs at the output.

6.3.2 Uncertainty-Aware Prediction Framework

The sparse nature of the glucose meter readings means that there is an associated uncertainty with the modelled glucose concentration levels over time. Incorporating this uncertainty can be beneficial in training a more robust neural network that is resilient to the various sources of noise. The approach we take in accomplishing this is to first align the posterior samples, $\mathbf{f} = \{f_1, f_2, ..., f_M\}$, generated from the GP with the corresponding time-series of the logs (meals, insulin, and exercise). These input sequences, $\mathbf{v} = \{v_1, v_2, ..., v_M\}$, are then mapped to their corresponding target label, y_i.

In this setting v_i is considered a random variable, therefore in order to appropriately train the neural network we minimise the expectation of the quantile loss with respect to the empirical average as shown in Equation 6.9.

$$\begin{aligned} w^* &= \underset{w}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{v}\sim\mathcal{N}(\mu,\mathbf{\Sigma};\theta)}[\ell(g(\mathbf{v}, w), y)], \\ &= \frac{1}{3}\sum_{q\in Q}\sum_{t=0}^{\tau}\sum_{m=1}^{M}\frac{q(y_t - g(\mathbf{v}, w))_+ + (1-q)(y_t - g(\mathbf{v}, w))_+}{M}, \end{aligned} \tag{6.9}$$

where w^* represents the optimal weights of the neural network parameters. The quantile outputs represent the predictions at the 2.5, 50, and 97.5 quantile.

6.4 Experiments

In this section we detail the experiments that are performed to demonstrate the benefit of the proposed approach. Firstly, we describe the dataset that is utilised in undertaking these experiments. We then introduce the baseline methods used to benchmark the different aspects of the proposed method. Lastly, we provide details on the multiple metrics and statistical analysis we use to comprehensively evaluate the performance of the proposed approach.

6.4.1 Dataset

OhioT1DM dataset. The OhioT1DM dataset [134] is introduced earlier as a publicly available dataset that comprises data collected from 12 participants who are individuals with T1D with glucose meters, wearable devices, insulin pumps, and a diary for self-reported logs. The individuals are observed in free-living conditions over a period of 8 weeks.

Data Pre-processing. Prior to training and testing, the data undergoes pre-processing to facilitate effective learning. The inputs considered for this models are the time of day, fingerstick measures, carbohydrate intake, insulin, and exercise. The time of day is first converted to seconds in the day and then a final signal through a cosine transform. The reported exercise is first transformed from the initial range of 1-10 to a binary representation denoting the presence (1) or absence (0) of exercise. The carbohydrate and insulin values are also scaled to within a 0-1 range. Finally, The glucose concentration values are scaled into the range [-1, 1]. This is first performed based on training data and then used to transform the test data.

Table 6.1: Training and testing data sizes and glucose meter readings (in brackets) for subjects in the OhioT1DM dataset

ID	Number of samples	
	Train	Test
540	11947 (414)	2884 (90)
544	10623 (191)	2704 (51)
552	9080 (180)	2352 (46)
559	10796 (169)	2514 (34)
563	12124 (531)	2570 (105)
567	10877 (406)	2377 (78)
570	10982 (241)	2745 (65)
575	11866 (253)	2590 (49)
584	12150 (135)	2653 (25)
588	12640 (558)	2791 (82)
591	10847 (411)	2760 (75)
596	10877 (166)	2731 (32)

6.4.2 Glucose Prediction Models

The following models are implemented to evaluate the different aspects of the proposed method:

LSTM. The first baseline method we consider is a deep learning model, an encoder-decoder architecture, that is based on a recurrent neural network (RNN) comprising LSTM cells. This deep learning model is included in order to demonstrate the benefit of a more flexible parametric model over the The irregularly-sampled measurements from the glucose meter are interpolated with a last value carry-forward rule.

LSTM-SHA. This baseline model is a deep learning model that features an encoder-decoder architecture based on a recurrent neural network comprising LSTM cells. The primary difference between this model and the LSTM model is the single head self-attention (SHA) layer. Furthermore, the difference between this model and the proposed approached is that the interpolation of the fingerstick measures is performed with a last-value carry forward approach as opposed to the GP-based approach.

GP-LSTM. Similarly, this approach is a two-stage model where the fingerstick measures are interpolated with the GP, and the mean of the posterior distribution serves as input to the LSTM neural network.

GP-LSTM-SHA. This approach is a two-stage model that uses interpolates fingerstick measures with a Gaussian process and performs the glucose forecasting with the LSTM-SHA model as explained in the Section 6.3. The main difference with the proposed approach is that this approach mainly takes the mean of the GP posterior distribution as input to the LSTM-SHA neural network.

UA-LSTM. This method is the proposed approach where the GP-LSTM is trained with an *uncertainty-aware* manner. The neural network is fed multiple samples from the posterior distribution of the GP model which map unto a single true output. In this study, we use 8 posterior samples for training all *uncertainty-aware* models as this provided the best validation performance.

Implementation Details

The models are developed with Python v3.6 using the GPy v1.9.9 and Tensorflow v1.15. The experiments were all performed using an NVIDIA GTX 1050.

6.4.3 Experimental Setup

Performance comparison of different nowcasting methods. In this experiment, we compare the performance of the proposed approach to the baseline methods in nowcasting the glucose concentrations levels in terms of the predictive and clinical accuracy. The models are evaluated using a held-out test set.

Adverse event detection compared to SMBG. This experiment is undertaken to determine the benefits of the proposed approach over the current SMBG approach. The proposed approach is evaluated based on the detection of hyperglycaemic events and temporal gain relative to SMBG.

6.4.4 Performance Metrics

We employ multiple criteria to comprehensively evaluate model performance in the areas of predictive accuracy, clinical accuracy, and detection time delay.

The proposed approach is geared towards nowcasting of glucose concentration levels. In order to evaluate the predictive accuracy, in terms of the magnitude of error from the reference CGM values in the dataset, we employ the standard metrics of RMSE and MAE.

To effectively evaluate the clinical utility of this approach, it would be necessary to consider the predicted glucose concentration levels as well as the rate and direction of change of the predicted glucose concentration values. The continuous glucose error grid analysis (CG-EGA), developed by Clarke *et al.* [186], is an augmentation of the original Clarke error grid analysis (EGA) [140] in order to better evaluate the significance of errors in diabetes management systems. This is achieved with the combination of two components, a point error grid (P-EGA) and a rate Error Grid (R-EGA) as shown below in Figure 6.4.

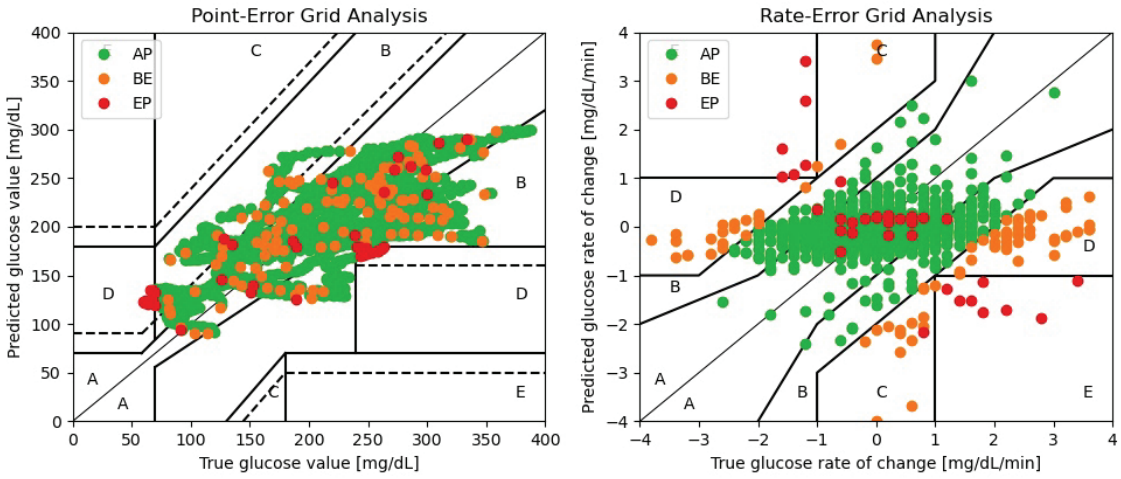

Figure 6.4: A visualisation of the error grid analysis (AP: Accurate Prediction; BE: Benign Error; EP: Erroneous Prediction) on a test set of an individual [4]. **Left:** The Point-Error Grid Analysis (P-EGA) measures the accuracy of the point predictions with the CGM reference. **Right:** The Rate-Error Grid Analysis (R-EGA) compares the rate of change of direction of the trajectory from -4mg/dL/min to +4mg/dL/min between reference and predicted rate of change.

The P-EGA is the same as the original EGA and measures the point accuracy of model predictions. On the other hand, the R-EGA analyses the clinical accuracy of the rate and the change of direction of the model predictions. Combining the P-EGA and the R-EGA is necessary to produce the final results, which are categorised

under one of the following labels: accurate prediction (AP), benign error (BE), and erroneous prediction (EP). The objective in this metric is to maximise AP and minimise BE and EP. These results are reported as a percentage over all the sample points as well as within the specific clinical regions - hypoglycaemia, euglycaemia, and hyperglycaemia.

Finally, for the comparison of eSMBG and SMBG methods we include an additional set of metrics. We employ precision, recall (sensitivity), and the F-score for evaluating the performance in detecting adverse glycaemic events. We also evaluate the proposed approach using a detection time delay. This is useful in determining the temporal gain in detecting the adverse glycaemic events for the proposed approach over the current SMBG method.

6.4.5 Statistical Analysis

For determining the statistical significance of differences between our proposed eSMBG approach and the other approaches, we first perform preliminary test for normality using the Shapiro-Wilk test. We use a paired t-test if normality is accepted, and a Wilcoxon signed-rank test if normality is rejected. The significance level is set at $p < .05$. For multiple pairwise comparisons, we adjust the significance level using Bonferroni correction.

6.5 Results

In this section, we present the summarised results of the experiments undertaken for this work along with observations relating to the model performance of the proposed approach.

Table 6.2: A comparison of the performance in terms of predictive accuracy and clinical accuracy between different models on the Ohio T1DM dataset (best result highlighted in **BOLD**)

Model	RMSE	MAE	CG-EGA		
			AP	BE	EP
LSTM	53.4±6.3*	42.0±4.5†	85.0±4.5	**6.4±1.7**	8.6±3.6
LSTM-SHA	53.6±6.8*	42.3±5.2†	84.2±4.4	7.7±1.9	8.1±3.7
GP-LSTM	51.5±6.3	40.5±4.8	**85.9±3.9**	6.6±1.8	**7.5±3.2**
GP-LSTM-SHA	53.0±7.2	41.8±5.6*	84.9±4.5	7.4±1.7	7.7±3.8
UA-LSTM	52.0±6.7*	40.9±5.0	85.1±4.6	6.8±1.9	8.1±3.5
eSMBG	**50.8±7.1**	**39.9±5.5**	85.6±4.3	6.9±1.6	**7.5±3.5**

$^{\frown}$ Statistical significance compared to eSMBG with p-value $< .01$ (*Paired t-test; †Wilcoxon)

6.5.1 Performance comparison of different methods.

Table 6.2 below provides a comparison of the results for the predictive accuracy and general clinical accuracy of the proposed and baseline approaches. The eSMBG approach (UA-LSTM-SHA) is shown to have the best predictive performance compared to the other approaches in terms of RMSE and MAE. eSMBG performance is significantly better than both LSTM and LSTM-SHA in terms of both MAE and RMSE. However, for the other GP-based approaches, eSMBG significantly outperforms UA-LSTM in terms of RMSE, and GP-LSTM-SHA in terms of MAE. In general, we also note that the predictive accuracy of the models are improved when the glucose concentration levels are modelled from the fingerstick measures using a Gaussian Process (GP) as opposed to the last value carry forward approach. Furthermore we see that although the uncertainty-aware training approach improves the general predictive accuracy for the proposed approach, this is not observed with GP-LSTM and UA-LSTM models.

We also observe the performance of the model performance in terms of clinical accuracy for the proposed approach and selected baselines. In this case the results are more nuanced, the proposed eSMBG approach (UA-LSTM-SHA) and the GP-LSTM show the lowest EP for the cohort tested. The GP-LSTM models also demonstrates the best performance of accurate predictions (AP), however this improvement is not significant relative to eSMBG. Similar to the case with predictive accuracy, we see

that the models where glucose concentration levels are modelled from fingerstick measures using a Gaussian Process regression approach exhibit lower erroneous predictions (EP) than models which used the last value carry forward approach.

Table 6.3 provides further analysis of the clinical accuracy in the specific glycaemic regions. As we note in the hypoglycaemia region, the performance of all models are less than 5% accurate predictions (AP) and the erroneous predictions (EP) are greater than 95% of the glucose predictions in the hypoglycaemia region. In the euglycaemia region, all the models demonstrate a high level of accurate predictions with AP >90% and low levels of erroneous predictions with EP <5%. In this case, the GP-LSTM demonstrates the best performance in the euglycaemia region. For our proposed approach, the uncertainty-aware training approach is shown to improve on GP-LSTM-SHA performance in the euglycaemia region.

Finally, clinical accuracy of models in the hyperglycaemia region is lower than in the euglycaemia region, but higher than in the hypoglycaemia region. The eSMBG and GP-LSTM-SHA models demonstrate the highest performance in the accurate prediction with 78.2% AP. However, GP-LSTM-SHA provides the lowest erroneous prediction with 12.3% EP. The eSMBG method shows the next best performance with the next lowest erroneous prediction with 13.0%.

Table 6.3: A comparison of the performance in terms of clinical accuracy between different models on the Ohio T1DM dataset in the different glycaemic regions (best result highlighted in **BOLD**)

Model	Hypoglycaemia			Euglycaemia			Hyperglycaemia		
	AP	BE	EP	AP	BE	EP	AP	BE	EP
LSTM	0.0±0.0	0.0±0.0	100.0±0.0	93.4±1.8	5.8±1.7	0.8±0.7	74.9±9.3	**8.5±2.9**	16.6±8.9
LSTM-SHA	2.4±7.5	0.0±0.0	97.6±7.5	91.4±2.3	6.8±1.8	1.7±1.3	76.0±8.6	10.2±3.1	13.7±8.4
GP-LSTM	0.2±0.4	0.2±0.4	99.7±0.6	**93.6±1.9**	**5.7±1.7**	**0.7±0.3**	78.0±8.6	8.8±3.2	13.2±8.5
GP-LSTM-SHA	**3.1±10.0**	0.2±0.4	**96.7±10.0**	91.9±2.5	6.8±1.8	1.3±1.0	**78.2±8.5**	9.6±2.7	**12.2±8.6**
UA-LSTM	0.0±0.1	0.0±0.1	99.9±0.3	92.4±2.4	6.2±1.9	1.3±0.9	77.2±8.9	8.8±3.2	14.0±8.8
eSMBG	0.1±0.2	0.0±0.0	99.9±0.3	92.7±1.7	6.4±1.4	1.0±0.4	**78.2±8.8**	8.8±2.8	13.0±8.5

In general, we note that the clinical accuracy in the hyperglycaemia region is generally better with GP-based models as these have lower EP than models where the glucose concentration is modelled with a last value carry forward approach.

6.5.2 Adverse event detection compared to SMBG

In Section 6.5.1, we compared the performance of the proposed approach against other forecasting algorithms. This demonstrated the predictive and clinical accuracy of the proposed model, particularly for the detection of hyperglycaemic events. Table 6.4 shows the improvement of the eSMBG approach over the current SMBG approach.

Table 6.4: A comparison of the performance in terms of event detection for hyperglycaemic events between SMBG and eSMBG. p-values calculated with Wilcoxon signed-rank test are underlined.

	Precision	Recall	F1	Detection Delay/mins
SMBG	0.59 ± 0.11	0.27 ± 0.12	0.36 ± 0.13	217 ± 71
eSMBG	0.60 ± 0.17	0.49 ± 0.18	0.52 ± 0.14	72 ± 67
% Δ	0.0%	+81.5%	+44.4%	−66.8%
p	<u>0.9</u>	<0.01	<0.01	<0.001

The eSMBG approach primarily leads to a significant increase (+81.5%, p <0.01) in the recall. As a result, this leads to a significant increase (+44.4%, p <0.01) in the F1 score although we do not see a significant change (+0.0%, p=0.9) in the precision. Furthermore, the eSMBG provides a significant reduction in the detection delay (-66.8%, p <0.001) of hyperglycaemic events with fingerstick measures.

6.6 Discussion

eSMBG is a deep learning framework that improves the capability of the self-monitoring of blood glucose to control glucose concentration levels. The strength of eSMBG lies in the uncertainty-aware learning approach and the presence of the attention components.

Attention-based deep learning models help provide a better performance as they model long-term dependencies [183]. The primary difference between the LSTM-based models and the attention-based models is the self-attention component. Typically, the LSTM-based models are able to model work by updating each successive LSTM cell with the state of the previous cell, hence the dependencies are predominantly local. However, self-attention can facilitate better connections between timesteps non-successive cells which may be difficult for LSTM networks to learn. From our results, we observe that the addition of an attention component can improve on the general predictive and clinical accuracy of the sequence-to-sequence model, however this seems to be contingent on a large amount of data being available as seen in the uncertainty-aware learning approach.

As noted earlier, the uncertainty-aware learning approach results from multiple signals being sampled from the posterior distribution, this is observed to mainly improve the overall predictive and clinical accuracy. This could be the result of the proposed approach having more layers, therefore requiring a large amount of data during the training phase. The posterior samples can be considered as a form of data augmentation, with the various posterior samples being perturbations on the mean sample to increase the number and diversity of data in the individual dataset [187, 188]. Typically, this is performed on image and text data to improve model performance in computer vision and natural language processing problems respectively [189, 190]. However, this approach could hamper performance when the model capacity is lower as observed in the case of the UA-LSTM model compared to the GP-LSTM model.

From the perspective of functionality, the use case of an eSMBG tool in diabetes management would be similar to the flash glucose monitor (FGM) [173]. The similarity between these tools is that they can provide more dynamic information such as trend information even with intermittent use. Furthermore, these tools do not require calibration, but are not particularly accurate for the region of hypoglycaemia and hence require support from additional fingerstick measures. However, results

reported in the literature [173, 170, 191] indicate that FGM currently exhibits better predictive accuracy relative to the demonstrated predictive accuracy for SMBG reported in this work. Another difference in this case is that eSMBG has no maximum sensor duration, unlike the FGM, and is only dependent on the availability of cheap test strips [26].

Ultimately, this approach faces limitations that impact model performance. Currently, the clinical guidance recommends particular moments when the fingerstick readings are taken with the glucose meter [7]. These guidelines are dictated by moments where subsequent actions would have an effective impact on glucose control. However, when modelling the glucose concentration levels from fingerstick measures, these points may not be the most optimal. In this case, future work could explore the optimal points for taking measurements to improve the modelling of glucose concentration levels, and subsequently predictive performance.

6.7 Conclusion

Self-monitoring of blood glucose with glucose meters remain the most accessible diabetes management tool available to individuals with Type 1 diabetes for aiding in glucose control. However, the sparse and irregular nature of fingerstick measures results in a much lower detection of adverse glycaemic events. We develop an uncertainty-aware deep learning model for providing continuous monitoring with short-term prediction of glucose concentration levels, which in turn improves detection of adverse glycaemic events, particularly hyperglycaemia. This approach of leveraging machine learning to improve existing tools can help narrow the gap in outcome disparities between the various CGM-based diabetes technology tools.

Chapter 7

Conclusion and Future Work

The advent of data engineering has laid the foundation for facilitating an improvement in the current diabetes technology tools that exist for individual living with with diabetes today. This thesis investigated the application of state-of-the-art machine learning methods towards enhancing the diabetes technology that enable individuals with diabetes to manage this chronic condition on a daily basis. Deep learning and multitask learning, specifically, allows the introduction of novel tools that provide more functionality of the main diabetes technology tools (i.e. SMBG, decision support, and artificial pancreas) and reduce the burden of managing diabetes.

7.1 Summary of Thesis Contributions

Chapter 3 introduced a deep learning framework for predicting the glucose prediction levels in mobile devices. The deep learning model proposed is based on a convolutional recurrent neural network (CRNN) that predicts the glucose concentration levels at both 30 minute and 60 minute prediction horizons. This makes use of information from wearable devices - continuous glucose monitors - and complementary information such as insulin and meals. This approach outperforms existing traditional approaches in the literature when evaluated on relevant metrics.

The application of the model is then further evaluated on a smartphone to demonstrate feasibility for on-device inference, and consequently, ambulatory monitoring and decision support on a resource-constrained device.

Chapter 4 extended the work presented in Chapter 3 to a multitask learning framework as a transfer learning strategy in developing personalised glucose prediction models. This work was motivated by the challenge of effectively learning from real-world population data in order to attain better personalised glucose prediction models when individual data is scarce. This perspective views each individual in the dataset as a unique task. In doing so, the multitask learning approach tackles the persistent issue of inter-individual variability that can hamper effective knowledge transfer and offers the opportunity to deploy such models earlier. Incorporating knowledge on glycaemic variability can help improve performance at long-term prediction horizons and minimise the effect of negative transfer, particularly at hypoglycaemia regions. Furthermore, in an application such as a predictive low-glucose suspend we show that this approach provides consistent performance with reduced requirements on the amount of training data required.

Chapter 5 introduced another multitask framework that is focused on moving the current hybrid artificial pancreas systems towards fully closed-loop insulin delivery. This is primarily tackled by substituting the meal announcement with a meal detection and estimation module as postprandial control is a significant hurdle to achieving fully closed-loop control. The problem is viewed from the perspective of outlier detection and the glucose trajectory estimation is performed with multiple quantile regression. In this setting, each task represents a pre-specified quantile to quantify the aleatoric uncertainty for the given glucose trajectory. A meal is therefore identified with a significant and persistent deviation from the expected trajectory.

On evaluation in an artificial pancreas system, this approach improves the glycaemic targets of the fully closed-loop setting and also improves the quality of glycaemic control over the hybrid approach.

Chapter 6 further demonstrates the utility of a deep learning framework in improving the functionality of diabetes technology tools. The proposed approach of an uncertainty-aware deep learning model allows more continuous monitoring of glucose concentration levels from SMBG tools. Furthermore, this approach increases the sensitivity (recall) of SMBG in detecting hyperglycaemic events earlier and without significantly degrading the precision.

7.2 Future Challenges and Perspectives

The methods introduced in this thesis work towards the ultimate goal of easing the burden of glucose control, however further opportunities and challenges still exist that can be addressed in future work.

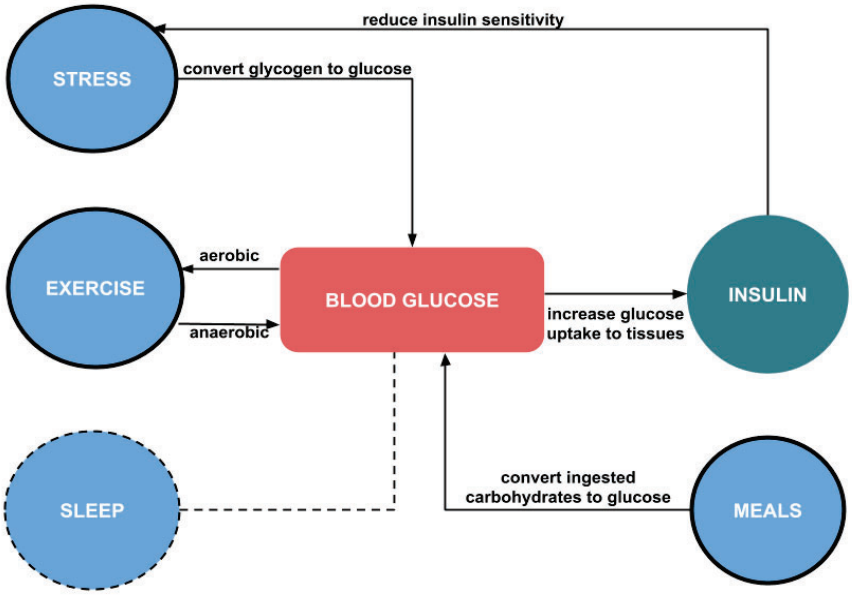

Figure 7.1: A simplified model of the external factors affecting blood glucose dynamics in T1DM population.

Incorporating physiological constraints into deep learning models. Although the improvement in diabetes technology tools is mainly fuelled by the performance improvements realised in data-driven methods, prior knowledge behind the physiological models such as Bergman minimal model, Hovorka model etc. re-

main important. This knowledge in the literature is developed from many years of experimental studies undertaken over the years to reveal the relations between the various factors i.e.*ceteris paribus* an increase in ingested meal increases glucose levels, and similarly, an increase in insulin leads to a decrease in glucose levels.

For the deep learning models, these relations are expected to be learned during the training process, however, this is demonstrated to not always be the case. For example, the work detailed in Chapter 5 requires a post-hoc verification of the physiological glucose-meal dynamics during meal estimation. From the viewpoint of safety and robustness, a next step would be explicitly incorporating prior physiological constraints into the deep learning models (physics-informed machine learning) such that the expected dynamics are baked into the learned representations of the models.

Incorporating more physiological signals into multitask learning models. The metabolic system is a complex system that is affected by other external factors beyond the primary considerations of meals and insulin [19, 15, 192]. Figure 7.1 illustrates that, for example, aerobic exercise can have the effect of reducing the blood glucose concentration level similar to insulin. Automatic detection of these events can further ease the burden of self-management on the individual with diabetes.

The existing work in this thesis has demonstrated the benefits of multitask learning and multitask architectures for developing diabetes technology tools. In Chapter 4 and 5 the multitask learning approach can facilitate effective information transfer between tasks and provide an architecture for efficient computation. In order to advance the utility of diabetes technology tools, these approaches can be exploited further in developing the next generation of deep learning models for diabetes management.

The deep learning models introduced in 3 and 5 can be further developed in the multitask learning framework to include detection of events such as exercise, sleep and stress. We posit that jointly learning these tasks as auxiliary tasks could also improve in the prediction of glucose values as these events would seem to be cor-

related. In terms of functionality, the detection of anaerobic exercise and/or stress could serve as a way to minimise the occurrence of false positives in our meal detection and estimation. Furthermore, as we see in Chapter 6, the performance of the model introduced is limited to providing trend information and forecasts during and shortly after glucose meter readings. This can be attributed to the rather sparse nature of the fingerstick measures. Given the possible correlation of these physiological signals that are more regularly sampled, we posit that a multitask Gaussian Process that models a sparse signal jointly with these more densely sampled signals could improve the prediction performance and allow continuous monitoring of glucose concentration levels, similar to minimally-invasive CGM devices.

7.3 Conclusion

The work in this thesis has detailed a number of approaches for that provide an avenue for improving the functionality of the current set of diabetes management tools available to individuals with diabetes. This can help expand the focus of the deep learning beyond glucose prediction, as seen with the applications such as that towards automatic meal detection. Given the increasing interest in deep learning, the introduction of complementary approaches such as multitask learning and uncertainty quantification in this thesis can lay the foundation for utilising deep learning to achieve the ultimate goal in developing the right tools for optimising glucose outcomes while minimising the burden of diabetes management.

Appendix A

List of Publications

Peer-Reviewed Journals:

J1. **J. Daniels**, P. Herrero and P. Georgiou, "Uncertainty-Aware Deep Learning for Enhancing Self Monitoring of Blood Glucose," *in preparation*, 2022. Appears in Chapter 6

J2. T. Zhu, L. Kuang, **J. Daniels**, P. Herrero, K. Li, and P. Georgiou, "IoMT-Enabled Real-time Blood Glucose Prediction with Deep Learning and Edge Computing," *IEEE Internet of Things Journal*, 2022.

J3. **J. Daniels**, P. Herrero and P. Georgiou, "A Deep Learning Framework for Automatic Meal Detection and Estimation in Artificial Pancreas Systems," *Sensors*, 2022. Appears in Chapter 5

J4. **J. Daniels**, P. Herrero and P. Georgiou, "A Multitask Learning Approach to Personalised Blood Glucose Prediction," *IEEE Journal of Biomedical and Health Informatics*, vol.26, no.1, pp.436-445, 2021. Appears in Chapter 4

J5. 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. Appears in Chapter 3

J6. P. Herrero, M. El-Sharkawy, **J. Daniels**, N. Jugnee, C. N. Uduku, M. Reddy, N. Oliver, and P. Georgiou, "The bio-inspired artificial pancreas for type 1 diabetes control in the home: System architecture and preliminary results," *Journal of diabetes science and technology*, vol.13, no.6, pp.1017-1025, 2019.

Conferences and Workshop Proceedings:

C1. **J. Daniels**, P. Herrero, P. Georgiou, "Automatic Meal Detection and Estimation using Neural Networks" in Diabetes Technology Therapeutics 23:A71-2, 2021. Appears in Chapter 5

C2. **J. Daniels**, T. Zhu, K. Li, C. Uduku, P. Herrero, N. Oliver, and P. Georgiou, "ARISES: An Advanced Clinical Decision Support Platform for the Management of Type 1 Diabetes," in Diabetes Technology Therapeutics, 22:A57, 2020 Appears in Chapter 3

C3. **J. Daniels** and P. Georgiou, "A Data-Driven Detection System for Predicting Stress Levels from Autonomic Signals," in 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), Oct. 2019, pp.1–4, iSSN: 2163-4025.

C4. M. El-Sharkawy, **J. Daniels**, P. Pesl, M. Reddy, N. Oliver, P. Herrero, and P. Georgiou, "A Portable Low-Power Platform for Ambulatory Closed Loop Control of Blood Glucose in Type 1 Diabetes," in 2018 IEEE International Symposium on Circuits and Systems (ISCAS), May 2018, pp. 1–5.

W1. **J. Daniels**, P. Herrero and P. Georgiou, "Personalised Glucose Prediction via Deep Multitask Networks," in Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data, KDH@ECAI 2020, Aug. 2020, vol. 2675 of CEUR Workshop Proceedings, pp. 110–114, CEUR-WS.org, 2020. Appears in Chapter 4

Appendix B

Supplementary Information - Chapter 3

B.1 Model architecture

Table B.1: A Table detailing the size and dimensions of layers in CRNN

Layer Description (layer)	Output Dimensions	No. of Parameters
Convolutional Layers (Batch×Steps×Channels)		
(1) 1×4 conv	128(1) × 24 × 8	104
max_pooling, size 2	128(1) × 12 × 8	–
(2) 1×4 conv	128(1) × 12 × 16	528
max_pooling, size 2	128(1) × 6 × 16	–
(3) 1×4 conv	128(1) × 6 × 32	2080
max_pooling	128(1) × 3 × 32	–
Recurrent Layer (Batch×Cells)		
(4) lstm	128(1) × 64	24832
Dense Layers (Batch×Units)		
(5) dense	128(1) × 256	16640
(6) dense	128(1) × 32	8224
(7) dense	128(1) × 1	33

Appendix C

Supplementary Information - Chapter 4

C.1 Clustering based on glycaemic variability

Glycaemic variability can represent a method for quantifying the degree of glycaemic excursions present in the individual for the observed period. This can be indicative of the risk hyperglycaemic or hypoglycaemia.

Table C.1: **Percentage time spent in hypoglycaemia for each individual in the training data:** The percentage time spent in hypoglycaemia is presented along with the associated glycaemic variability (represented by the coefficient of variation (CV)) for every individual.

ID	Glycaemic Variability (CV) (%)	Percentage time spent in hypogylcaemia (%)
540	40	7.1
544	36	1.5
552	37	3.7
559	42	4.1
563	34	2.6
567	40	6.7
570	33	2.0
575	43	8.7
584	34	0.9
588	31	1.0
591	37	3.9
596	33	2.1

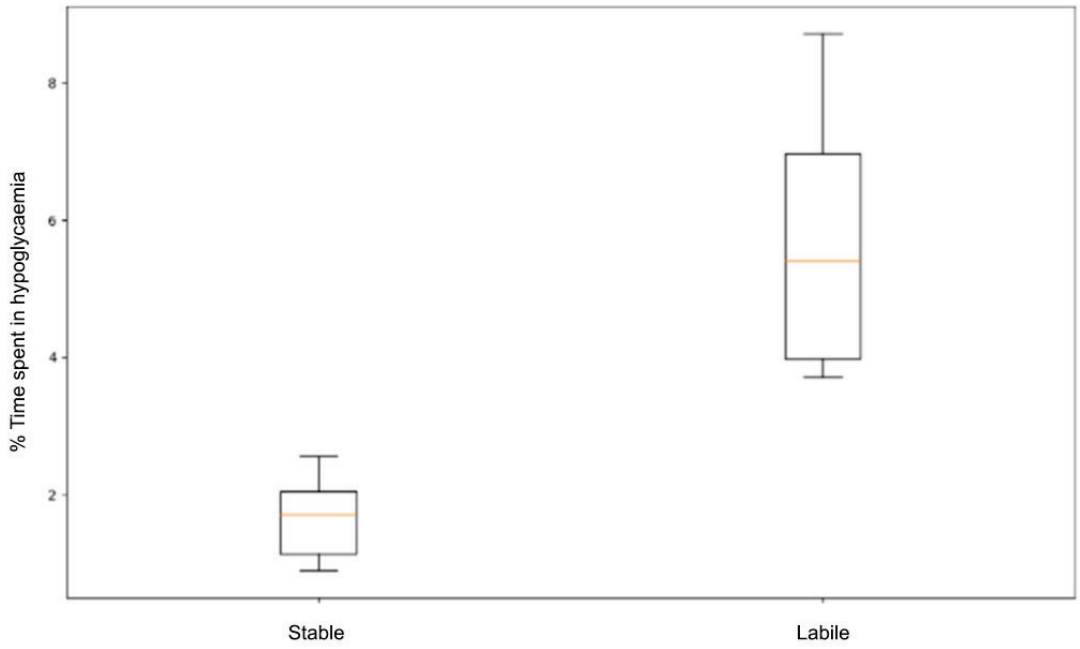

Figure C.1: **Comparison of percentage time in hypoglycaemia between stable and labile groups:** Based on a 36% threshold for separating individuals in stable (CV *leq* 36%) and labile (CV > 36%) groups, a Mann-Whitney U test shows a significant difference in percentage time in hypoglycaemia (p = .003).

Mean Percentage time spent in hypoglycaemia for stable group = 1.7 ± 0.6 %

Mean Percentage time spent in hypoglycaemia for labile group = 5.7 ± 1.9 %

C.2 Bayesian hyperparameter optimisation

To select hyperparameters for the glucose prediction models, we perform a Bayesian hyperparameter optimization algorithm with Tree of Parzen estimators using CometML [193]. The search space and optimal hyperparameters are shown in Table C.2.

Table C.2: Hyperparameter search space and configuration

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

C.3 Extended quantitative results on model performance

Table C.3: **Model performance in terms of RMSE and MAE in the Ohio T1DM test dataset:** The predictive accuracy for each model, in terms of RMSE and MAE, in the test dataset at all considered prediction horizons. The performance is reported for each glycaemic region.

Prediction Horizon (mins)	Model	HYPOGLYCAEMIA BG $\leq$ 70mg/dL		EUGLYCAEMIA 70 mg/dL<BG<180mg/dL		HYPERGLYCAEMIA BG $\geq$ 180mg/dL	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
30	MTL	15.8 ± 4.3	12.6 ± 4.1	16.5 ± 2.2	11.9 ± 1.7	22.7 ± 3.9	16.1 ± 2.5
	MTL-GV	14.7 ± 4.1	11.4 ± 4.1	16.5 ± 2.0	11.8 ± 1.6	22.8 ± 4.0	16.1 ± 2.6
	TL	15.0 ± 3.3	12.1 ± 3.3	16.7 ± 2.1	12.0 ± 1.7	23.5 ± 4.1	16.6 ± 2.6
	STL	17.5 ± 3.7	14.9 ± 3.9	18.4 ± 2.3	13.4 ± 2.0	24.7 ± 4.3	17.8 ± 3.1
	SVR	16.8 ± 5.4	14.1 ± 5.4	17.1 ± 2.3	12.3 ± 1.9	22.8 ± 4.8	16.0 ± 2.6
45	MTL	25.8 ± 5.1	21.7 ± 5.6	21.9 ± 2.4	16.0 ± 2.1	31.0 ± 5.0	22.6 ± 3.7
	MTL-GV	26.7 ± 5.5	22.6 ± 5.2	23.1 ± 3.0	17.1 ± 2.5	30.6 ± 4.2	22.4 ± 3.0
	TL	24.6 ± 4.5	20.1 ± 5.3	23.0 ± 2.8	16.8 ± 2.3	32.6 ± 5.3	23.6 ± 3.8
	STL	25.5 ± 6.0	21.4 ± 6.7	23.5 ± 3.0	17.3 ± 2.5	32.6 ± 5.5	24.0 ± 4.0
	SVR	27.1 ± 8.4	23.4 ± 8.6	23.5 ± 3.0	17.3 ± 2.7	31.6 ± 6.0	22.8 ± 3.9
60	MTL	36.4 ± 5.9	30.9 ± 6.5	27.1 ± 3.7	20.2 ± 2.8	39.4 ± 6.4	29.6 ± 4.9
	MTL-GV	34.2 ± 6.3	28.1 ± 6.9	27.8 ± 3.7	20.5 ± 2.9	40.0 ± 6.3	30.1 ± 5.0
	TL	37.1 ± 6.2	31.6 ± 6.2	28.6 ± 3.6	21.3 ± 2.7	40.4 ± 6.1	30.5 ± 4.7
	STL	40.9 ± 8.6	35.5 ± 9.5	29.2 ± 3.9	21.8 ± 3.1	41.7 ± 6.7	31.8 ± 5.7
	SVR	39.6 ± 11.8	34.4 ± 12.1	28.6 ± 3.8	21.4 ± 3.4	38.7 ± 6.6	28.6 ± 4.9
90	MTL	56.5 ± 11.2	51.3 ± 11.6	35.4 ± 4.1	27.1 ± 3.2	50.1 ± 8.1	38.6 ± 7.1
	MTL-GV	51.8 ± 9.1	46.5 ± 9.5	34.0 ± 3.5	25.8 ± 2.7	52.9 ± 7.5	41.3 ± 7.0
	TL	55.0 ± 13.1	49.4 ± 12.9	35.8 ± 4.7	27.1 ± 3.5	54.3 ± 7.8	42.6 ± 6.7
	STL	58.5 ± 11.0	53.2 ± 11.1	35.5 ± 4.1	27.3 ± 3.7	54.1 ± 9.6	43.0 ± 8.9
	SVR	63.7 ± 18.3	58.4 ± 18.4	37.0 ± 6.6	28.8 ± 5.9	49.6 ± 8.9	38.2 ± 7.6
120	MTL	72.9 ± 12.9	68.9 ± 13.3	41.5 ± 5.8	32.4 ± 4.7	57.5 ± 9.5	45.8 ± 8.3
	MTL-GV	68.7 ± 12.2	64.6 ± 11.7	39.3 ± 5.0	30.7 ± 3.8	58.5 ± 8.5	46.6 ± 7.5
	TL	71.6 ± 12.9	67.8 ± 11.8	39.7 ± 6.1	30.8 ± 4.7	62.5 ± 9.1	51.4 ± 7.9
	STL	70.9 ± 14.9	66.7 ± 14.4	40.4 ± 5.9	31.3 ± 4.6	61.6 ± 10.6	50.3 ± 9.6
	SVR	71.9 ± 12.4	68.9 ± 12.1	36.6 ± 5.0	29.1 ± 4.1	63.0 ± 10.7	52.3 ± 10.3

Table C.4: **Model performance in terms of EGA in the Ohio T1DM test dataset:** The EGA performance for each model, showing the percentage of CGM samples in each zone (A-E) is reported for the test dataset at all considered prediction horizons. This also shows the combined performance at Zone A and B (Zone A+B).

Prediction Horizon	Model	EGA					
		Zone A+B	Zone A	Zone B	Zone C	Zone D	Zone E
(mins)	MTL	99.1 ± 0.7	89.8 ± 3.8	9.2 ± 3.2	0.0 ± 0.0	0.9 ± 0.6	0.0 ± 0.0
	MTL-GV	99.2 ± 0.5	90.0 ± 3.6	9.2 ± 3.2	0.0 ± 0.0	0.8 ± 0.5	0.0 ± 0.0
30	TL	99.1 ± 0.6	89.6 ± 3.5	9.5 ± 3.1	0.0 ± 0.0	0.9 ± 0.6	0.0 ± 0.0
	STL	98.6 ± 1.4	87.0 ± 4.9	11.6 ± 3.9	0.0 ± 0.0	1.4 ± 1.3	0.0 ± 0.0
	SVR	98.9 ± 0.8	89.4 ± 4.1	9.5 ± 3.4	0.0 ± 0.1	1.0 ± 0.8	0.0 ± 0.0
	MTL	97.9 ± 1.5	81.2 ± 6.3	16.7 ± 5.1	0.0 ± 0.0	2.1 ± 1.5	0.0 ± 0.0
	MTL-GV	97.5 ± 2.0	79.4 ± 7.4	18.2 ± 5.8	0.1 ± 0.1	2.4 ± 2.0	0.0 ± 0.0
45	TL	98.1 ± 1.3	80.2 ± 6.2	17.9 ± 5.1	0.1 ± 0.1	1.9 ± 1.3	0.0 ± 0.0
	STL	97.8 ± 1.7	78.9 ± 7.5	18.9 ± 6.1	0.1 ± 0.1	2.1 ± 1.6	0.0 ± 0.0
	SVR	97.7 ± 1.7	78.9 ± 7.3	18.8 ± 5.9	0.1 ± 0.1	2.2 ± 1.6	0.0 ± 0.1
	MTL	96.8 ± 2.1	72.3 ± 8.2	24.5 ± 6.5	0.2 ± 0.2	3.0 ± 2.0	0.0 ± 0.0
	MTL-GV	97.1 ± 2.0	72.2 ± 7.9	24.9 ± 6.2	0.2 ± 0.2	2.7 ± 1.8	0.0 ± 0.0
60	TL	96.6 ± 2.3	70.2 ± 7.9	26.5 ± 6.0	0.3 ± 0.2	3.1 ± 2.1	0.0 ± 0.0
	STL	96.2 ± 2.8	68.8 ± 9.2	27.4 ± 6.8	0.3 ± 0.3	3.5 ± 2.6	0.0 ± 0.0
	SVR	96.6 ± 2.2	70.7 ± 9.5	26.0 ± 7.4	0.3 ± 0.3	3.0 ± 2.1	0.1 ± 0.0
	MTL	95.0 ± 3.0	58.9 ± 8.3	36.0 ± 5.8	0.5 ± 0.3	4.5 ± 2.8	0.0 ± 0.1
	MTL-GV	95.0 ± 2.9	59.6 ± 8.0	35.4 ± 5.7	0.5 ± 0.4	4.5 ± 2.7	0.0 ± 0.1
90	TL	94.6 ± 2.8	57.6 ± 8.4	37.0 ± 6.1	0.6 ± 0.5	4.8 ± 2.5	0.1 ± 0.1
	STL	94.5 ± 3.0	57.8 ± 8.0	36.7 ± 5.8	0.7 ± 0.4	4.8 ± 2.9	0.1 ± 0.1
	SVR	94.9 ± 3.0	57.6 ± 9.5	37.3 ± 7.6	0.7 ± 0.6	4.3 ± 2.7	0.1 ± 0.1
	MTL	93.7 ± 3.0	51.8 ± 7.3	41.9 ± 5.0	0.9 ± 0.6	5.2 ± 2.9	0.2 ± 0.2
	MTL-GV	93.8 ± 2.8	52.3 ± 7.5	41.6 ± 5.3	0.7 ± 0.5	5.4 ± 2.7	0.1 ± 0.1
120	TL	92.8 ± 2.9	50.0 ± 7.4	42.9 ± 5.6	0.9 ± 0.7	6.1 ± 2.8	0.2 ± 0.2
	STL	93.1 ± 3.3	50.5 ± 8.5	42.6 ± 6.1	1.0 ± 0.7	5.8 ± 3.0	0.2 ± 0.2
	SVR	93.1 ± 3.6	50.9 ± 8.9	42.2 ± 6.5	0.4 ± 0.4	6.4 ± 3.5	0.1 ± 0.1

Appendix D

Supplementary Information - Chapter 5

D.1 Meal detection and estimation algorithm

In the meal detection and estimation algorithm, the multitask quantile regression model which forms the basis of this algorithm requires 28 time steps to start working. The decoder we set N = 2 to ensure that the flag is only activated once the majority (at least 15 min out of 20 min) of estimated samples are below the CGM trajectory. Δt is set at 20 min. The maximum meal estimate, m_{MAX}, is set to 90g at a time to mitigate risk of overestimation; additional successive meal estimates. During meal estimation appropriate glucose dynamics is verified done by tracking the MAE of glucose as the meal size is incremented. The meal estimate is discarded if successive increments in meal size lead to an increase in MAE as this violates the expected dynamics.

D.2 Training and validation of multitask quantile regression model

The model architecture is primarily a multitask encoder-decoder architecture based on LSTM recurrent neural networks. The LSTM layers for both the encoder layer and decoder layer consist of 64 cells. The selected optimiser is the Adam optimiser and at the pre-training stage the learning rate is 1×10^{-3}, which is then reduced to 1×10^{-4} at the fine-tuning stage. The batch size for both training stages is 128. We set the number of epochs to 100 and implement early stopping with a patience of 20 epochs to terminate training when validation loss is no longer improving ($\Delta\mathcal{L}_{min} = 1 \times 10^{-4}$). The models are developed with Python 3.6 and Keras v2.2.2 [113] and trained using a NVIDIA GTX 1050.

Table D.1: **Performance metrics for neural network on validation set:** The parameters of the meal detection and estimation framework are selected based on performance of the model on the validation set. This is based on last 20% of the training data in order to attain the predictive accuracy and prediction interval necessary.

ID	Metric		
	RMSE	MAE	PI
1	5.8	4.0	95.3
2	5.9	4.1	94.5
3	8.0	4.6	94.7
4	5.7	4.0	94.9
5	7.7	4.3	93.6
6	5.7	4.1	94.9
7	6.7	4.8	94.7
8	7.2	4.3	93.7
9	5.0	3.7	97.3
10	8.4	4.3	95.4
Average	**6.6 ± 1.1**	**4.2 ± 0.3**	**94.9 ± 1.0**

Appendix E

Supplementary Information - Chapter 6

E.1 Bayesian hyperparameter optimisation

To select hyperparameters for the glucose prediction models, we perform a Bayesian hyperparameter optimization algorithm with Tree of Parzen estimators using CometML [193]. The search space and optimal hyperparameters are shown in Table E.1.

Table E.1: Hyperparameter search space and configuration

Hyperparameter	Value Range	LSTM	LSTM-SHA
Number of cells	{8,16,32,64}	16	8
Hidden layer size	{8,16,32,64}	8	8
Learning rate	$\{1 \times 10^{-5}, ..., 1 \times 10^{-2}\}$	9×10^{-4}	3×10^{-3}
Max Gradient Norm	$\{1 \times 10^{-2}, ..., 10\}$	2×10^{-1}	7×10^{-1}
Dropout rate	{0.10,..., 0.90}	0.13	0.27
Batch size	{64, 128, 256, 512, 1024}		
	FF	256	64
	GP	256	64
	UA	256	1024

⁀ FF = Last Value Carry Forward; GP = Gaussian Process; UA = Uncertainty-Aware Training

Bibliography

[1] R. Yeo and M. Sawdon, "Hormonal control of metabolism: regulation of plasma glucose," *Anaesthesia & Intensive Care Medicine*, vol. 14, pp. 296–300, July 2013.

[2] "Dexcom continuous glucose monitoring." https://www.dexcom.com/en-NZ. Accessed: 2018-11-27.

[3] G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Calibration of Minimally Invasive Continuous Glucose Monitoring Sensors: State-of-The-Art and Current Perspectives," *Biosensors*, vol. 8, p. 24, Mar. 2018.

[4] M. De Bois, "Cg-ega python implementation," 2019. doi: 10.5281/zenodo.3459485.

[5] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 138, pp. 271–281, Apr. 2018.

[6] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, Nov. 2019.

[7] A. D. Association, "7. Diabetes Technology: Standards of Medical Care in Diabetes—2020," *Diabetes Care*, vol. 43, pp. S77–S88, Jan. 2020. Publisher: American Diabetes Association Section: Position Statements.

[8] M. J. Davies, D. A. D'Alessio, J. Fradkin, W. N. Kernan, C. Mathieu, G. Mingrone, P. Rossing, A. Tsapas, D. J. Wexler, and J. B. Buse, "Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)," *Diabetes Care*, p. dci180033, Sept. 2018.

[9] S. F. Clarke and J. R. Foster, "A history of blood glucose meters and their role in self-monitoring of diabetes mellitus," *British Journal of Biomedical Science*, vol. 69, pp. 83–93, Jan. 2012.

[10] H. S. Mariani, B. T. Layden, and G. Aleppo, "Continuous Glucose Monitoring: A Perspective on Its Past, Present, and Future Applications for Diabetes Management," *Clinical Diabetes*, vol. 35, pp. 60–65, Jan. 2017.

[11] G. Reach, "Can Continuous Glucose Monitoring Be Used for the Treatment of Diabetes," p. 6.

[12] Giacomo Cappon, Giada Acciaroli, Martina Vettoretti, Andrea Facchinetti, and Giovanni Sparacino, "Wearable Continuous Glucose Monitoring Sensors: A Revolution in Diabetes Treatment," *Electronics*, vol. 6, p. 65, Sept. 2017.

[13] S. R. Colberg, R. J. Sigal, J. E. Yardley, M. C. Riddell, D. W. Dunstan, P. C. Dempsey, E. S. Horton, K. Castorino, and D. F. Tate, "Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association," *Diabetes Care*, vol. 39, pp. 2065–2079, Nov. 2016.

[14] N. K. Brockman, R. J. Sigal, G. P. Kenny, M. C. Riddell, B. A. Perkins, and J. E. Yardley, "Sex-Related Differences in Blood Glucose Responses to Resistance Exercise in Adults With Type 1 Diabetes: A Secondary Data Analysis," *Canadian Journal of Diabetes*, vol. 44, pp. 267–273.e1, Apr. 2020.

[15] R. S. Surwit, M. S. Schneider, and M. N. Feinglos, "Stress and Diabetes Mellitus," *Diabetes Care*, vol. 15, pp. 1413–1422, Oct. 1992.

[16] J. Waterhouse, "Homeostatic control mechanisms," *Anaesthesia & Intensive Care Medicine*, vol. 14, pp. 291–295, July 2013.

[17] National Diabetes Data Group, "Classification and Diagnosis of Diabetes Mellitus and Other Categories of Glucose Intolerance," *Diabetes*, vol. 28, pp. 1039–1057, Dec. 1979.

[18] American Diabetes Association Workgroup on Hypoglycemia, "Defining and Reporting Hypoglycemia in Diabetes: A report from the American Diabetes Association Workgroup on Hypoglycemia," *Diabetes Care*, vol. 28, pp. 1245–1249, May 2005.

[19] M. C. Riddell, I. W. Gallen, C. E. Smart, C. E. Taplin, P. Adolfsson, A. N. Lumb, A. Kowalski, R. Rabasa-Lhoret, R. J. McCrimmon, C. Hume, F. Annan, P. A. Fournier, C. Graham, B. Bode, P. Galassetti, T. W. Jones, I. S. Millán, T. Heise, A. L. Peters, A. Petz, and L. M. Laffel, "Exercise management in type 1 diabetes: a consensus statement," *The Lancet Diabetes & Endocrinology*, vol. 5, pp. 377–390, May 2017.

[20] P. Galassetti and M. C. Riddell, "Exercise and type 1 diabetes (T1DM)," *Comprehensive Physiology*, vol. 3, pp. 1309–1336, July 2013.

[21] S. J. Kelly and M. Ismail, "Stress and Type 2 Diabetes: A Review of How Stress Contributes to the Development of Type 2 Diabetes," *Annual Review of Public Health*, vol. 36, pp. 441–462, Mar. 2015.

[22] J. Sturt, K. Dennick, M. Due-Christensen, and K. McCarthy, "The detection and management of diabetes distress in people with type 1 diabetes," *Current Diabetes Reports*, vol. 15, p. 101, Nov. 2015.

[23] M. E. Hilliard, J. P. Yi-Frazier, D. Hessler, A. M. Butler, B. J. Anderson, and S. Jaser, "Stress and A1c Among People with Diabetes Across the Lifespan," *Current Diabetes Reports*, vol. 16, p. 67, June 2016.

[24] I. Kyrou and C. Tsigos, "Stress hormones: physiological stress and regulation of metabolism," *Current Opinion in Pharmacology*, vol. 9, pp. 787–793, Dec. 2009.

[25] M. L. Marcovecchio and F. Chiarelli, "The Effects of Acute and Chronic Stress on Diabetes Control," *Science Signaling*, vol. 5, pp. pt10–pt10, Oct. 2012.

[26] J. Hönes, P. Müller, and N. Surridge, "The Technology Behind Glucose Meters: Test Strips," *Diabetes Technology & Therapeutics*, vol. 10, pp. S–10, May 2008. Num Pages: S-26 Publisher: Mary Ann Liebert, Inc., publishers.

[27] K. Tonyushkina and J. H. Nichols, "Glucose Meters: A Review of Technical Challenges to Obtaining Accurate Results," *Journal of Diabetes Science and Technology*, vol. 3, pp. 971–980, July 2009. Publisher: SAGE Publications Inc.

[28] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Wearable Continuous Glucose Monitoring Sensors: A Revolution in Diabetes Treatment," *Electronics*, vol. 6, p. 65, Sept. 2017. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[29] T. J. D. R. F. C. G. M. S. Group, "Continuous Glucose Monitoring and Intensive Treatment of Type 1 Diabetes," July 2009.

[30] J. P. New, R. Ajjan, A. F. H. Pfeiffer, and G. Freckmann, "Continuous glucose monitoring in people with diabetes: the randomized controlled Glucose Level Awareness in Diabetes Study (GLADIS)," *Diabetic Medicine: A Journal of the British Diabetic Association*, vol. 32, pp. 609–617, May 2015.

[31] M. Lind, W. Polonsky, I. B. Hirsch, T. Heise, J. Bolinder, S. Dahlqvist, E. Schwarz, A. F. Ólafsdóttir, A. Frid, H. Wedel, E. Ahlén, T. Nyström,

and J. Hellman, "Continuous Glucose Monitoring vs Conventional Therapy for Glycemic Control in Adults With Type 1 Diabetes Treated With Multiple Daily Insulin Injections: The GOLD Randomized Clinical Trial," *JAMA*, vol. 317, pp. 379–387, Jan. 2017.

[32] R. M. Bergenstal, D. C. Klonoff, S. K. Garg, B. W. Bode, M. Meredith, R. H. Slover, A. J. Ahmann, J. B. Welsh, S. W. Lee, and F. R. Kaufman, "Threshold-Based Insulin-Pump Interruption for Reduction of Hypoglycemia," *New England Journal of Medicine*, vol. 369, pp. 224–232, July 2013.

[33] A. Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, and G. Hartvigsen, "Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes," *Artificial Intelligence in Medicine*, vol. 98, pp. 109–134, July 2019.

[34] C. Cobelli, E. Renard, and B. Kovatchev, "Artificial Pancreas: Past, Present, Future," *Diabetes*, vol. 60, pp. 2672–2682, Nov. 2011.

[35] C. K. Boughton and R. Hovorka, "Advances in artificial pancreas systems," *Science Translational Medicine*, vol. 11, Mar. 2019. Publisher: American Association for the Advancement of Science Section: Focus.

[36] R. Ramli, M. Reddy, and N. Oliver, "Artificial Pancreas: Current Progress and Future Outlook in the Treatment of Type 1 Diabetes," *Drugs*, vol. 79, pp. 1089–1101, July 2019.

[37] N. Allen and A. Gupta, "Current Diabetes Technology: Striving for the Artificial Pancreas," *Diagnostics*, vol. 9, p. 31, Mar. 2019. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[38] J. Kesavadev, S. Srinivasan, B. Saboo, M. Krishna B, and G. Krishnan, "The Do-It-Yourself Artificial Pancreas: A Comprehensive Review," *Diabetes Therapy*, vol. 11, pp. 1217–1235, June 2020.

[39] E. Salzsieder, A. Rutscher, and U. Fischer, "KADIS: A computer-aided decision support, system to improve the quality of metabolic care in diabetes," in *1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 6, pp. 2284–2285, Oct. 1992.

[40] M. Berger, M. Jaffe, P. Devine, and D. Rodbard, "Computer Programs to Assist the Physician in the Analysis of Self-Monitored Blood Glucose Data," *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 52–57, Nov. 1988.

[41] M. P. Berger, R. A. Gelfand, and P. L. Miller, "Combining statistical, rule-based, and physiologic model-based methods to assist in the management of diabetes mellitus," *Computers and Biomedical Research*, vol. 23, pp. 346–357, Aug. 1990.

[42] G. Cappon, A. Facchinetti, G. Sparacino, P. Georgiou, and P. Herrero, "Classification of Postprandial Glycemic Status with Application to Insulin Dosing in Type 1 Diabetes—An In Silico Proof-of-Concept," *Sensors*, vol. 19, p. 3168, Jan. 2019.

[43] G. Cappon, M. Vettoretti, F. Marturano, A. Facchinetti, and G. Sparacino, "A Neural-Network-Based Approach to Personalize Insulin Bolus Calculation Using Continuous Glucose Monitoring," *Journal of Diabetes Science and Technology*, vol. 12, pp. 265–272, Mar. 2018.

[44] G. Noaro, G. Cappon, M. Vettoretti, G. Sparacino, S. D. Favero, and A. Facchinetti, "Machine-Learning Based Model to Improve Insulin Bolus Calculation in Type 1 Diabetes Therapy," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 247–255, Jan. 2021. Conference Name: IEEE Transactions on Biomedical Engineering.

[45] J. Burdick, H. P. Chase, R. H. Slover, K. Knievel, L. Scrimgeour, A. K. Maniatis, and G. J. Klingensmith, "Missed insulin meal boluses and elevated

hemoglobin A1c levels in children receiving insulin pump therapy," *Pediatrics*, vol. 113, pp. e221–224, Mar. 2004.

[46] L. M. Norlander, S. Anderson, C. J. Levy, L. Ekhlaspour, D. W. Lam, L. Hsu, S. E. Loebner, S. J. Ogyaadu, G. O'malley, C. M. Levister, M. D. Breton, and B. Buckingham, "Late and Missed Meal Boluses with Multiple Daily Insulin Injections," *Diabetes*, vol. 67, July 2018. Publisher: American Diabetes Association Section: General Poster Session.

[47] S. Samadi, M. Rashid, K. Turksoy, J. Feng, I. Hajizadeh, N. Hobbs, C. Lazaro, M. Sevil, E. Littlejohn, and A. Cinar, "Automatic Detection and Estimation of Unannounced Meals for Multivariable Artificial Pancreas System," *Diabetes Technology & Therapeutics*, vol. 20, pp. 235–246, Mar. 2018.

[48] J. Xie and Q. Wang, "A Variable State Dimension Approach to Meal Detection and Meal Size Estimation: In Silico Evaluation Through Basal-Bolus Insulin Therapy for Type 1 Diabetes," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 1249–1260, June 2017.

[49] C. M. Ramkissoon, P. Herrero, J. Bondia, and J. Vehi, "Unannounced Meals in the Artificial Pancreas: Detection Using Continuous Glucose Monitoring," *Sensors*, vol. 18, p. 884, Mar. 2018. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[50] Z. Mahmoudi, K. Nørgaard, N. K. Poulsen, H. Madsen, and J. B. Jørgensen, "Fault and meal detection by redundant continuous glucose monitors and the unscented Kalman filter," *Biomedical Signal Processing and Control*, vol. 38, pp. 86–99, Sept. 2017.

[51] N. S. Tyler and P. G. Jacobs, "Artificial Intelligence in Decision Support Systems for Type 1 Diabetes," *Sensors*, vol. 20, p. 3214, Jan. 2020. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[52] M. Gadaleta, A. Facchinetti, E. Grisan, and M. Rossi, "Prediction of Adverse Glycemic Events From Continuous Glucose Monitoring Signal," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, pp. 650–659, Mar. 2019.

[53] A. Güemes, G. Cappon, B. Hernandez, M. Reddy, N. Oliver, P. Georgiou, and P. Herrero, "Predicting Quality of Overnight Glycaemic Control in Type 1 Diabetes Using Binary Classifiers," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 1439–1446, May 2020. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[54] A. Parcerisas, I. Contreras, A. Delecourt, A. Bertachi, A. Beneyto, I. Conget, C. Viñals, M. Giménez, and J. Vehi, "A Machine Learning Approach to Minimize Nocturnal Hypoglycemic Events in Type 1 Diabetic Patients under Multiple Doses of Insulin," *Sensors*, vol. 22, p. 1665, Jan. 2022. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[55] 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, June 2019.

[56] R. Reddy, N. Resalat, L. M. Wilson, J. R. Castle, J. El Youssef, and P. G. Jacobs, "Prediction of Hypoglycemia During Aerobic Exercise in Adults With Type 1 Diabetes," *Journal of Diabetes Science and Technology*, vol. 13, pp. 919–927, Sept. 2019.

[57] 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.

[58] R. N. Bergman, L. S. Phillips, and C. Cobelli, "Physiologic evaluation of factors controlling glucose tolerance in man: measurement of insulin sensitivity and beta-cell glucose sensitivity from the response to intravenous glucose.,"

The Journal of Clinical Investigation, vol. 68, pp. 1456–1467, Dec. 1981. Publisher: American Society for Clinical Investigation.

[59] 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, and M. E. Wilinska, "Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes," *Physiological Measurement*, vol. 25, pp. 905–920, July 2004. Publisher: IOP Publishing.

[60] S. Samadi, K. Turksoy, I. Hajizadeh, J. Feng, M. Sevil, and A. Cinar, "Meal Detection and Carbohydrate Estimation Using Continuous Glucose Sensor Data," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 619–627, May 2017.

[61] R. Nimri, E. Dassau, T. Segall, I. Muller, N. Bratina, O. Kordonouri, R. Bello, T. Biester, K. Dovc, A. Tenenbaum, A. Brener, M. Šimunović, S. D. Sakka, M. Nevo Shenker, C. G. Passone, I. Rutigliano, D. Tinti, C. Bonura, S. Caiulo, A. Ruszala, B. Piccini, D. Giri, R. Stein, I. Rabbone, P. Bruzzi, J. Omladič, C. Steele, G. Beccuti, M. Yackobovitch-Gavan, T. Battelino, T. Danne, E. Atlas, and M. Phillip, "Adjusting insulin doses in patients with type 1 diabetes who use insulin pump and continuous glucose monitoring: Variations among countries and physicians," *Diabetes, Obesity & Metabolism*, vol. 20, pp. 2458–2466, Oct. 2018.

[62] C. Zhao, E. Dassau, L. Jovanovič, H. C. Zisser, F. J. Doyle, and D. E. Seborg, "Predicting Subcutaneous Glucose Concentration Using a Latent-Variable-Based Statistical Method for Type 1 Diabetes Mellitus," *Journal of Diabetes Science and Technology*, vol. 6, pp. 617–633, May 2012.

[63] D. A. Finan, F. J. Doyle III, C. C. Palerm, W. C. Bevier, H. C. Zisser, L. Jovanovič, and D. E. Seborg, "Experimental evaluation of a recursive model identification technique for type 1 diabetes," *Journal of diabetes science and technology*, vol. 3, no. 5, pp. 1192–1202, 2009.

[64] 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, pp. 81–88, Jan. 2010.

[65] E. I. Georga, V. C. Protopappas, D. Ardigò, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, "Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, pp. 71–81, Jan. 2013.

[66] M. Akbari and R. Chunara, "Using Contextual Information to Improve Blood Glucose Prediction," *arXiv:1909.01735 [cs, q-bio, stat]*, Aug. 2019. arXiv: 1909.01735.

[67] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "A predictive model of subcutaneous glucose concentration in type 1 diabetes based on Random Forests," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2889–2892, Aug. 2012. ISSN: 1558-4615.

[68] D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalová, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav, and E. Segal, "Personalized Nutrition by Prediction of Glycemic Responses," *Cell*, vol. 163, pp. 1079–1094, Nov. 2015.

[69] P. Herrero, P. Pesl, M. Reddy, N. Oliver, P. Georgiou, and C. Toumazou, "Advanced Insulin Bolus Advisor Based on Run-To-Run Control and Case-Based Reasoning," *IEEE journal of biomedical and health informatics*, vol. 19, pp. 1087–1096, May 2015.

[70] E. M. Aiello, C. Toffanin, M. Messori, C. Cobelli, and L. Magni, "Postprandial Glucose Regulation via KNN Meal Classification in Type 1 Diabetes," *IEEE Control Systems Letters*, vol. 3, pp. 230–235, Apr. 2019.

[71] E. Dassau, B. W. Bequette, B. A. Buckingham, and F. J. Doyle, "Detection of a Meal Using Continuous Glucose Monitoring: Implications for an artificial -cell," *Diabetes Care*, vol. 31, pp. 295–300, Feb. 2008.

[72] R. A. Harvey, E. Dassau, H. Zisser, D. E. Seborg, and F. J. Doyle, "Design of the Glucose Rate Increase Detector: A Meal Detection Module for the Health Monitoring System," *Journal of Diabetes Science and Technology*, vol. 8, pp. 307–320, Mar. 2014. Publisher: SAGE Publications Inc.

[73] J. Garcia-Tirado, J. L. Diaz, R. Esquivel-Zuniga, C. L. K. Koravi, J. P. Corbett, M. Dawson, C. Wakeman, C. L. Barnett, M. C. Oliveri, H. Myers, K. Krauthause, M. D. Breton, and M. D. DeBoer, "Advanced Closed-Loop Control System Improves Postprandial Glycemic Control Compared With a Hybrid Closed-Loop System Following Unannounced Meal," *Diabetes Care*, vol. 44, pp. 2379–2387, Oct. 2021. Publisher: American Diabetes Association Section: Emerging Technologies: Data Systems and Devices.

[74] M. M. Ahmadi and G. A. Jullien, "A wireless-implantable microsystem for continuous blood glucose monitoring," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, pp. 169–180, Jun. 2009.

[75] A. Facchinetti, "Continuous glucose monitoring sensors: Past, present and future algorithmic challenges," *Sensors*, vol. 16, no. 12, 2016.

[76] S. Oviedo, J. Veh, 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. 2833, 2017.

[77] M. Vettoretti, A. Facchinetti, G. Sparacino, and C. Cobelli, "Type 1 diabetes patient decision simulator for in silico testing safety and effectiveness of insulin treatments," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2018.

[78] P. Pesl, P. Herrero, M. Reddy, M. Xenou, N. Oliver, D. Johnston, C. Toumazou, and P. Georgiou, "An advanced bolus calculator for type 1 diabetes: System architecture and usability results," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, pp. 11–17, Jan 2016.

[79] A. G. Karegowda, M. Jayaram, and A. Manjunath, "Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients," *International Journal of Engineering and Advanced Technology*, vol. 1, no. 3, pp. 2249 – 8958, 2012.

[80] E. I. Georga, V. C. Protopappas, D. Ardig, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, pp. 71–81, Jan 2013.

[81] K. Yan and D. Zhang, "Blood glucose prediction by breath analysis system with feature selection and model fusion," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6406–6409, Aug. 2014.

[82] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[83] K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Systems with Applications*, vol. 34, no. 1, pp. 482 – 487, 2008.

[84] 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.

[85] C. Zecchin, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1550–1560, Jun. 2012.

[86] K. Plis, R. Bunescu, C. Marling, J. Shubrook, and F. Schwartz, "A machine learning approach to predicting blood glucose levels for diabetes management," in *Modern Artificial Intelligence for Health Analytics Papers from the AAAI-14*, 2014.

[87] H. Mhaskar, S. Pereverzyev, and M. van der Walt, "A deep learning approach to diabetic blood glucose prediction," *https://arxiv.org/abs/1707.05828*, 2017.

[88] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction," in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data*, (Stockholm, Sweden), July 2018.

[89] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pp. 675–678, 2014.

[90] G. Litjens, C. I. Snchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. H. van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports*, vol. 6, p. 26286, May 2016.

[91] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, pp. 1236–1246, 05 2017.

[92] 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*.

[93] Y. Bengio, "Deep learning of representations: Looking forward," in *Statistical Language and Speech Processing*, (Berlin, Heidelberg), pp. 1–37, Springer Berlin Heidelberg, 2013.

[94] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015.

[95] K. Li, A. Javer, E. Keaveny, and A. Brown, "Recurrent neural networks with interpretable cells predict and classify worm behaviour," in *Workshop on Worm's Neural Information Processing (WNIP) in NIPS*, (CA, USA), Dec. 2017.

[96] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8827–8836, Jun. 2018.

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

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

[99] C. Zhao, E. Dassau, L. Jovanovic, H. C. Zisser, I. Francis J. Doyle, and D. E. Seborg, "Predicting subcutaneous glucose concentration using a latent-

variable-based statistical method for type 1 diabetes mellitus," *Journal of Diabetes Science and Technology*, vol. 6, no. 3, pp. 617–633, 2012.

[100] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The uva/padova type 1 diabetes simulator," *Jounral of Diabetes Sci Technol.*, vol. 8, pp. 26–34, Jan. 2014.

[101] P. Herrero, J. Bondia, O. Adewuyi, P. Pesl, M. El-Sharkawy, M. Reddy, C. Toumazou, N. Oliver, and P. Georgiou, "Enhancing automatic closed-loop glucose control in type 1 diabetes with an adaptive meal bolus calculator–in silico evaluation under intra-day variability," *Computer Methods and Programs in Biomedicine*, vol. 146, pp. 125–131, 2017.

[102] M. Reddy, P. Pesl, M. Xenou, C. Toumazou, D. Johnston, P. Georgiou, P. Herrero, and N. Oliver, "Clinical safety and feasibility of the advanced bolus calculator for type 1 diabetes based on case-based reasoning: A 6-week nonrandomized single-arm pilot study," *Diabetes Technology & Therapeutics*, vol. 18, no. 8, pp. 487–493, 2016.

[103] 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, pp. 931–937, May 2007.

[104] S. G. Mougiakakou, A. Prountzou, D. Iliopoulou, K. S. Nikita, A. Vazeou, and C. S. Bartsocas, "Neural network based glucose - insulin metabolism models for children with type 1 diabetes," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3545–3548, Aug. 2006.

[105] R. Atanassov, P. Bose, M. Couture, A. Maheshwari, P. Morin, M. Paquette, M. Smid, and S. Wuhrer, "Algorithms for optimal outlier removal," *Journal of Discrete Algorithms*, vol. 7, no. 2, pp. 239 – 248, 2009. Selected papers from the 2nd Algorithms and Complexity in Durham Workshop ACiD 2006.

[106] H. Shum, K. Ikeuchi, and R. Reddy, *Principal Component Analysis with Missing Data and Its Application to Polyhedral Object Modeling*, pp. 3–39. Boston, MA: Springer US, 2001.

[107] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy k-means clustering method," in *Rough Sets and Current Trends in Computing*, (Berlin, Heidelberg), pp. 573–579, Springer Berlin Heidelberg, 2004.

[108] L. Lekha and S. M, "Real-time non-invasive detection and classification of diabetes using modified convolution neural network," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1630–1636, 2018.

[109] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[110] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE Int. Conf. on Acou., Spe. and Sig. Proc.*, p. 6645, 2013.

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

[112] S. Ruder, "An overview of gradient descent optimization algorithms," 2016.

[113] F. Chollet *et al.*, "Keras." https://github.com/fchollet/keras, 2015.

[114] International Hypoglycaemia Study Group, "Glucose Concentrations of Less Than 3.0 mmol/L (54 mg/dL) Should Be Reported in Clinical Trials: A Joint Position Statement of the American Diabetes Association and the European Association for the Study of Diabetes: Table 1," *Diabetes Care*, vol. 40, pp. 155–157, Jan. 2017.

[115] X. Liu, I. Jankovic, and J. H. Chen, "Predicting Inpatient Glucose Levels and Insulin Dosing by Machine Learning on Electronic Health Records," Mar. 2020. Pages: 2020.03.02.20029017.

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

[117] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep Learning for Diabetes: A Systematic Review," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[118] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?," *arXiv:1608.08614 [cs]*, Dec. 2016. arXiv: 1608.08614.

[119] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice), pp. 843–852, IEEE, Oct. 2017.

[120] A. Ceriello, L. Monnier, and D. Owens, "Glycaemic variability in diabetes: clinical and therapeutic implications," *The Lancet Diabetes & Endocrinology*, Aug. 2018.

[121] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *arXiv:1911.02685 [cs, stat]*, June 2020. arXiv: 1911.02685.

[122] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji, "Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, (Sydney, NSW, Australia), pp. 1475–1484, ACM Press, 2015.

[123] T. Killian, G. Konidaris, and F. Doshi-Velez, "Transfer Learning Across Patient Variations with Hidden Parameter Markov Decision Processes," *arXiv:1612.00475 [cs, stat]*, Dec. 2016. arXiv: 1612.00475.

[124] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health," *IEEE Transactions on Affective Computing*, pp. 1–1, 2017.

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

[126] A. R. Bhimireddy, P. Sinha, B. Oluwalade, J. W. Gichoya, and S. Purkayastha, "Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks," in *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data, KDH@ECAI 2020* (K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, eds.), vol. 2675 of *CEUR Workshop Proceedings*, pp. 125–130, CEUR-WS.org, 2020.

[127] H. Rubin-Falcone, I. Fox, and J. Wiens, "Deep residual time-series forecasting: Application to blood glucose prediction," in *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020* (K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, eds.), vol. 2675 of *CEUR Workshop Proceedings*, pp. 105–109, CEUR-WS.org, 2020.

[128] H. Hameed and S. Kleinberg, "Investigating potentials and pitfalls of knowledge distillation across datasets for blood glucose forecasting," in *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020* (K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, eds.), vol. 2675 of *CEUR Workshop Proceedings*, pp. 85–89, CEUR-WS.org, 2020.

[129] W. Gu, Z. Zhou, Y. Zhou, M. He, H. Zou, and L. Zhang, "Predicting Blood Glucose Dynamics with Multi-time-series Deep Learning," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems - SenSys '17*, (Delft, Netherlands), pp. 1–2, ACM Press, 2017.

[130] M. De Bois, M. A. E. Yacoubi, and M. Ammi, "Adversarial Multi-Source Transfer Learning in Healthcare: Application to Glucose Prediction for Diabetic People," *arXiv:2006.15940 [cs, eess]*, June 2020. arXiv: 2006.15940.

[131] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv:1503.02531 [cs, stat]*, Mar. 2015. arXiv: 1503.02531.

[132] T. Kushner, M. D. Breton, and S. Sankaranarayanan, "Multi-Hour Blood Glucose Prediction in Type 1 Diabetes: A Patient-Specific Approach Using Shallow Neural Network Models," *Diabetes Technology & Therapeutics*, p. dia.2020.0061, Oct. 2020.

[133] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, pp. 41–75, July 1997.

[134] C. Marling and R. C. Bunescu, "The ohiot1dm dataset for blood glucose level prediction: Update 2020," in *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020* (K. Bach, R. C. Bunescu, C. Marling, and N. Wiratunga, eds.), vol. 2675 of *CEUR Workshop Proceedings*, pp. 71–74, CEUR-WS.org, 2020.

[135] L. Monnier, C. Colette, A. Wojtusciszyn, S. Dejager, E. Renard, N. Molinari, and D. R. Owens, "Toward Defining the Threshold Between Low and High Glucose Variability in Diabetes," *Diabetes Care*, vol. 40, pp. 832–838, July 2017.

[136] 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, pp. 603–613, Feb. 2020. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[137] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, and R. Ranganath, "Opportunities in Machine Learning for Healthcare," *arXiv:1806.00388 [cs, stat]*, June 2018. arXiv: 1806.00388.

[138] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[139] 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, pp. 603–613, Feb. 2020. Conference Name: IEEE Journal of Biomedical and Health Informatics.

[140] W. L. Clarke, "The original Clarke Error Grid Analysis (EGA)," *Diabetes Technology & Therapeutics*, vol. 7, pp. 776–779, Oct. 2005.

[141] G. P. Forlenza, Z. Li, B. A. Buckingham, J. E. Pinsker, E. Cengiz, R. P. Wadwa, L. Ekhlaspour, M. M. Church, S. A. Weinzimer, E. Jost, T. Marcal, C. Andre, L. Carria, V. Swanson, J. W. Lum, C. Kollman, W. Woodall, and R. W. Beck, "Predictive Low-Glucose Suspend Reduces Hypoglycemia in Adults, Adolescents, and Children With Type 1 Diabetes in an At-Home Randomized Crossover Study: Results of the PROLOG Trial," *Diabetes Care*, vol. 41, pp. 2155–2161, Oct. 2018.

[142] W. A. Pruett, J. S. Clemmer, and R. L. Hester, "Physiological Modeling and Simulation—Validation, Credibility, and Application," *Annual Re-*

view of Biomedical Engineering, vol. 22, no. 1, pp. 185–206, 2020. _eprint: https://doi.org/10.1146/annurev-bioeng-082219-051740.

[143] E. Dassau, H. Zisser, R. A. Harvey, M. W. Percival, B. Grosman, W. Bevier, E. Atlas, S. Miller, R. Nimri, L. Jovanovič, and F. J. Doyle, "Clinical Evaluation of a Personalized Artificial Pancreas," *Diabetes Care*, vol. 36, pp. 801–809, Apr. 2013. Publisher: American Diabetes Association Section: Original Research.

[144] M. Reddy, P. Herrero, M. E. Sharkawy, P. Pesl, N. Jugnee, D. Pavitt, I. F. Godsland, G. Alberti, C. Toumazou, D. G. Johnston, P. Georgiou, and N. S. Oliver, "Metabolic Control With the Bio-inspired Artificial Pancreas in Adults With Type 1 Diabetes: A 24-Hour Randomized Controlled Crossover Study," *Journal of Diabetes Science and Technology*, vol. 10, pp. 405–413, Nov. 2015.

[145] S. Robinson, R. S. Newson, B. Liao, T. Kennedy-Martin, and T. Battelino, "Missed and mistimed insulin doses in people with diabetes: a systematic literature review," *Diabetes Technology & Therapeutics*, July 2021. Publisher: Mary Ann Liebert, Inc., publishers.

[146] D. M. Hessler, L. Fisher, W. H. Polonsky, U. Masharani, L. A. Strycker, A. L. Peters, I. Blumer, and V. Bowyer, "Diabetes distress is linked with worsening diabetes management over time in adults with Type 1 diabetes," *Diabetic Medicine: A Journal of the British Diabetic Association*, vol. 34, pp. 1228–1234, Sept. 2017.

[147] K. Smythe, M. Saw, M. Mak, and V. W. Wong, "Carbohydrate knowledge, lifestyle and insulin: an observational study of their association with glycaemic control in adults with type 1 diabetes," *Journal of Human Nutrition and Dietetics: The Official Journal of the British Dietetic Association*, vol. 31, pp. 597–602, Oct. 2018.

[148] K. L. Joiner, M. L. Holland, and M. Grey, "Stressful Life Events in Young Adults With Type 1 Diabetes in the U.S. T1D Exchange Clinic Registry," *Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing*, vol. 50, pp. 676–686, Nov. 2018.

[149] W. V. Tamborlane, K. M. Pfeiffer, M. Brod, A. Nikolajsen, A. Sandberg, A. L. Peters, and M. Van Name, "Understanding bolus insulin dose timing: the characteristics and experiences of people with diabetes who take bolus insulin," *Current Medical Research and Opinion*, vol. 33, pp. 639–645, Apr. 2017.

[150] K. A. Datye, C. T. Boyle, J. Simmons, D. J. Moore, S. S. Jaser, N. Sheanon, J. M. Kittelsrud, S. E. Woerner, and K. M. Miller, "Timing of Meal Insulin and Its Relation to Adherence to Therapy in Type 1 Diabetes," *Journal of Diabetes Science and Technology*, vol. 12, pp. 349–355, Mar. 2018. Publisher: SAGE Publications Inc.

[151] L. M. Norlander, E. T. Nykaza, B. Arbiter, B. A. Buckingham, and R. Lal, "931-P: Hyperglycemia following Early vs. Late Meal Boluses," *Diabetes*, vol. 68, June 2019. Publisher: American Diabetes Association Section: Poster Presentations: Clinical Diabetes/Therapeutics.

[152] A. Peters, M. A. Van Name, B. Thorsted, J. S. Piltoft, and W. V. Tamborlane, "Postprandial Dosing of Bolus Insulin in Patients with type 1 Diabetes: a Cross-sectional Study Using Data From the T1d Exchange Registry," *Endocrine Practice*, vol. 23, pp. 1201–1209, Oct. 2017.

[153] A. S. Brazeau, H. Mircescu, K. Desjardins, C. Leroux, I. Strychar, J. M. Ekoé, and R. Rabasa-Lhoret, "Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes," *Diabetes Research and Clinical Practice*, vol. 99, pp. 19–23, Jan. 2013.

[154] L. T. Meade and W. E. Rushton, "Accuracy of Carbohydrate Counting in Adults," *Clinical Diabetes*, vol. 34, pp. 142–147, July 2016. Publisher: American Diabetes Association Section: Feature Articles.

[155] F. Cameron, G. Niemeyer, and B. A. Buckingham, "Probabilistic Evolving Meal Detection and Estimation of Meal Total Glucose Appearance," *Journal of Diabetes Science and Technology*, vol. 3, pp. 1022–1030, Sept. 2009. Publisher: SAGE Publications Inc.

[156] M. Zheng, B. Ni, and S. Kleinberg, "Automated meal detection from continuous glucose monitor data through simulation and explanation," *Journal of the American Medical Informatics Association*, vol. 26, pp. 1592–1599, Dec. 2019.

[157] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks.," in *NIPS* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 3104–3112, 2014.

[158] 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 *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 706–712, July 2019. ISSN: 1557-170X.

[159] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[160] N. Tagasovska and D. Lopez-Paz, "Single-Model Uncertainties for Deep Learning," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[161] L. Magni, D. M. Raimondo, C. D. Man, M. Breton, S. Patek, G. D. Nicolao, C. Cobelli, and B. P. Kovatchev, "Evaluating the Efficacy of Closed-Loop Glucose Regulation via Control-Variability Grid Analysis," *Journal of diabetes science and technology (Online)*, vol. 2, pp. 630–635, July 2008.

[162] C. Liu, J. Vehí, P. Avari, M. Reddy, N. Oliver, P. Georgiou, and P. Herrero, "Long-Term Glucose Forecasting Using a Physiological Model and Deconvolution of the Continuous Glucose Monitoring Signal," *Sensors*, vol. 19, pp. 4338 – 4357, Jan. 2019. Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[163] A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," *Biostatistics*.

[164] A. Kendall and Y. Gal, "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?," *arXiv:1703.04977 [cs]*, Oct. 2017. arXiv: 1703.04977.

[165] D. M. Nathan and f. t. D. R. Group, "The Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study at 30 Years: Overview," *Diabetes Care*, vol. 37, pp. 9–16, Jan. 2014. Publisher: American Diabetes Association Section: DCCT/EDIC 30th Anniversary Summary Findings.

[166] I. Mamkin, S. Ten, S. Bhandari, and N. Ramchandani, "Real-Time Continuous Glucose Monitoring in the Clinical Setting: The Good, the Bad, and the Practical," *Journal of Diabetes Science and Technology*, vol. 2, pp. 882–889, Sept. 2008. Publisher: SAGE Publications Inc.

[167] K. D. Barnard, T. Wysocki, J. M. Allen, D. Elleri, H. Thabit, L. Leelarathna, A. Gulati, M. Nodale, D. B. Dunger, T. Tinati, and R. Hovorka, "Closing the loop overnight at home setting: psychosocial impact for adolescents with type 1 diabetes and their parents," *BMJ Open Diabetes Research and Care*, vol. 2, p. e000025, Apr. 2014. Publisher: BMJ Specialist Journals Section: Clinical care/education/nutrition/psychosocial research.

[168] J. C. Wong, N. C. Foster, D. M. Maahs, D. Raghinaru, R. M. Bergenstal, A. J. Ahmann, A. L. Peters, B. W. Bode, G. Aleppo, I. B. Hirsch, L. Kleis,

H. P. Chase, S. N. DuBose, K. M. Miller, R. W. Beck, S. Adi, and for the T1D Exchange Clinic Network, "Real-Time Continuous Glucose Monitoring Among Participants in the T1D Exchange Clinic Registry," *Diabetes Care*, vol. 37, pp. 2702–2709, Sept. 2014.

[169] E. Toschi and H. Wolpert, "Utility of Continuous Glucose Monitoring in Type 1 and Type 2 Diabetes," *Endocrinology and Metabolism Clinics of North America*, vol. 45, pp. 895–904, Dec. 2016.

[170] R. A. Ajjan, M. H. Cummings, P. Jennings, L. Leelarathna, G. Rayman, and E. G. Wilmot, "Accuracy of flash glucose monitoring and continuous glucose monitoring technologies: Implications for clinical practice," *Diabetes and Vascular Disease Research*, vol. 15, pp. 175–184, May 2018. Publisher: SAGE Publications.

[171] Y. C. Kudva, A. J. Ahmann, R. M. Bergenstal, J. R. Gavin, III, D. F. Kruger, L. K. Midyett, E. Miller, and D. R. Harris, "Approach to Using Trend Arrows in the FreeStyle Libre Flash Glucose Monitoring Systems in Adults," *Journal of the Endocrine Society*, vol. 2, pp. 1320–1337, Dec. 2018.

[172] K. M. Miller, R. W. Beck, R. M. Bergenstal, R. S. Goland, M. J. Haller, J. B. McGill, H. Rodriguez, J. H. Simmons, I. B. Hirsch, and T1D Exchange Clinic Network, "Evidence of a strong association between frequency of self-monitoring of blood glucose and hemoglobin A1c levels in T1D exchange clinic registry participants," *Diabetes Care*, vol. 36, pp. 2009–2014, July 2013.

[173] G. Mancini, M. G. Berioli, E. Santi, F. Rogari, G. Toni, G. Tascini, R. Crispoldi, G. Ceccarini, and S. Esposito, "Flash Glucose Monitoring: A Review of the Literature with a Special Focus on Type 1 Diabetes," *Nutrients*, vol. 10, p. 992, Aug. 2018. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[174] 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, June 2017.

[175] C. C. Palerm, H. Zisser, W. C. Bevier, L. Jovanovič, and F. J. Doyle, III, "Prandial Insulin Dosing Using Run-to-Run Control: Application of clinical data and medical expertise to define a suitable performance metric," *Diabetes Care*, vol. 30, pp. 1131–1136, May 2007.

[176] M. Cescon, S. Deshpande, R. Nimri, F. J. D. III, and E. Dassau, "Using Iterative Learning for Insulin Dosage Optimization in Multiple-Daily-Injections Therapy for People With Type 1 Diabetes," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 482–491, Feb. 2021. Conference Name: IEEE Transactions on Biomedical Engineering.

[177] W. Gu, "Non-intrusive blood glucose monitor by multi-task deep learning: PhD forum abstract," in *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks - IPSN '17*, (Pittsburgh, Pennsylvania), pp. 249–250, ACM Press, 2017.

[178] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep Interpretable Early Warning System for the Detection of Clinical Deterioration," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 437–446, Feb. 2020.

[179] J. Futoma, S. Hariharan, and K. Heller, "Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier," *arXiv:1706.04152 [stat]*, June 2017. arXiv: 1706.04152.

[180] M. Moor, M. Horn, B. Rieck, D. Roqueiro, and K. Borgwardt, "Early Recognition of Sepsis with Gaussian Process Temporal Convolutional Networks and Dynamic Time Warping," *arXiv:1902.01659 [cs, stat]*, Oct. 2020. arXiv: 1902.01659.

[181] M. Rosnati and V. Fortuin, "MGP-AttTCN: An Interpretable Machine Learning Model for the Prediction of Sepsis," *arXiv:1909.12637 [cs, stat]*, Sept. 2019. arXiv: 1909.12637 version: 1.

[182] I. Urteaga, T. Bertin, T. M. Hardy, D. J. Albers, and N. Elhadad, "Multi-Task Gaussian Processes and Dilated Convolutional Networks for Reconstruction of Reproductive Hormonal Dynamics," p. 24, 2019.

[183] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), pp. 6000–6010, Curran Associates Inc., Dec. 2017.

[184] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," *arXiv:1912.09363 [cs, stat]*, Dec. 2019. arXiv: 1912.09363.

[185] C. E. Rasmussen, "Gaussian Processes in Machine Learning," in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (O. Bousquet, U. von Luxburg, and G. Rätsch, eds.), Lecture Notes in Computer Science, pp. 63–71, Berlin, Heidelberg: Springer, 2004.

[186] B. P. Kovatchev, L. A. Gonder-Frederick, D. J. Cox, and W. L. Clarke, "Evaluating the accuracy of continuous glucose-monitoring sensors: continuous glucose-error grid analysis illustrated by TheraSense Freestyle Navigator data," *Diabetes Care*, vol. 27, pp. 1922–1928, Aug. 2004.

[187] T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid, "A Bayesian Data Augmentation Approach for Learning Deep Models," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[188] T. DeVries and G. W. Taylor, "Dataset Augmentation in Feature Space," *arXiv:1702.05538 [cs, stat]*, Feb. 2017. arXiv: 1702.05538.

[189] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A Survey of Data Augmentation Approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, (Online), pp. 968–988, Association for Computational Linguistics, Aug. 2021.

[190] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, p. 60, July 2019.

[191] L. Leelarathna and E. G. Wilmot, "Flash forward: a review of flash glucose monitoring," *Diabetic Medicine*, vol. 35, no. 4, pp. 472–482, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/dme.13584.

[192] M. B. Davidson, M. D. Harris, F. H. Ziel, and C. S. Rosenberg, "Suppression of Sleep-Induced Growth Hormone Secretion by Anticholinergic Agent Abolishes Dawn Phenomenon," vol. 37, p. 6, 1988.

[193] "Comet.ML." https://www.comet.ml/. Accessed: 2021-2-3.