Aalto University
School of Science
Master's Programme in Life Science Technologies

Viljami Raiski

Exploring Latent Structure in Continuous Glucose Monitor Data

Master's Thesis
Espoo, May 23, 2021

Supervisor: Professor Juho Rousu, Aalto University
Advisors: Sandor Szedmak Ph.D, Aalto University
Liisa Hakaste MD, Ph.D, University of Helsinki

Aalto University
School of Science
Master's Programme in Life Science Technologies

ABSTRACT OF
MASTER'S THESIS

Author: Viljami Raiski

Title:
Exploring Latent Structure in Continuous Glucose Monitor Data

Date:	May 23, 2021	**Pages:**	65 + 5
Major:	Bioinformatics and Digital Health	**Code:**	SCI3092
Supervisor:	Professor Juho Rousu, Aalto University		
Advisors:	Sandor Szedmak Ph.D, Aalto University Liisa Hakaste MD, Ph.D, University of Helsinki		

Diabetes mellitus is a growing problem in both developing and developed countries, affecting the lives of over 422 million people worldwide. The traditional way of diagnosing and detecting diabetes has been using single-time point measurements. However, the arrival of continuous glucose monitoring (CGM) systems has enabled access to the blood glucose time series.

The blood glucose time series provided by CGM technology have been utilised in numerous use-cases such as forecasting blood glucose values and classifying diabetes via supervised learning methods, detecting abnormal patterns via unsupervised learning methods, and designing blood glucose control algorithms using reinforcement learning methods.

This thesis explores the applicability of supervised learning methods for capturing the latent structure in an individual's blood glucose dynamics. Furthermore, we experiment with the usefulness of the captured latent structure in two use-cases, classification between subjects diagnosed with type 2 diabetes (T2D) and prediabetes, and predicting blood sugar surges to a hyperglycemic level.

Using CGM data from 62 subjects diagnosed with either T2D or prediabetes, we used Linear Regression and Kernel Regression to model blood glucose time series. After fitting the regression models, we extracted the regression coefficients and used them as a glucose profile vector characterising the latent structure in the blood glucose data.

The extracted blood glucose profile was shown to improve performance in glycemic variability based T2D/prediabetes classification. In the hyperglycemia prediction, the glucose profile did not have a notable impact on the performance.

Thesis results suggest that regression coefficients could capture relevant information about an individual's blood glucose dynamics, but further research and clinical validation is required due to the small sample size (N=62) and possible inaccuracies in the CGM technology. Furthermore, the achieved results should be compared with current medically validated methods for assessing blood glucose dynamics.

Keywords: continuous glucose monitor, blood glucose, diabetes, prediabetes, supervised learning

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Master's Programme in Life Science Technologies

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Viljami Raiski

Työn nimi:
Piilevien rakenteiden tutkiminen jatkuvasta glukoosinseurantadatasta

Päiväys: 23. toukokuuta 2021 **Sivumäärä:** 65 + 5

Pääaine: Bioinformatics and Digital Health **Koodi:** SCI3092

Valvoja: Professori Juho Rousu, Aalto-yliopisto

Ohjaajat: Tohtori Sandor Szedmak, Aalto-yliopisto
Lääketieteen tohtori Liisa Hakaste, Helsingin yliopisto

Diabetes mellitus on kasvava maailmanlaajuinen ongelma sekä kehittyneissä, että kehittyvissä maissa. Maailmanlaajuisesti diabetes mellitus vaikuttaa arviolta 422 miljoonan ihmisen elämään. Nykyinen tapa tunnistaa diabetes mellitus perustuu pääosin hetkittäisen verensokeriarvojen tarkasteluun, esimerkiksi oraaliseen glukoosikokeeseen ja paastoverensokeriin. Jatkuvan glukoosimonitoroinnin (CGM) myötä on kyetty tutkimaan myös yksilöiden verensokerin aikasarjoja.

CGM-teknologian tuottamaa verensokeriaikasarjaa on hyödynnetty lukuisissa koneoppimisen sovelluksissa. Ohjatun oppimisen sovelluksia ovat olleet veren glukoositasojen ennustaminen, hyper- ja hypoglykemian ennustaminen ja diabeteksen diagnosointi. Ohjaamattomassa oppimisessa tutkimus on keskittynyt verensokeriaikasarjojen piilevän rakenteen tutkimukseen. Vahvistusoppimisessa tutkimus on keskittynyt verensokerin säätelyalgoritmien kehitykseen.

Tässä diplomityössä tutkitaan ohjattujen oppimismenetelmien soveltuvuutta piilevän rakenteen tutkimiseen yksilön verensokerin dynamiikassa. Kokeellisessa osuudessa tutkimme piilevän rakenteen hyödyllisyyttä kahdessa käyttötapauksessa, tyypin 2 diabeetikoiden (T2D) ja esidiabeetikoiden luokittelussa ja hetkittäisen korkean verensokerin ennustuksessa.

Hyödyntäen T2D tai esidiabetesdiagnoosin saaneiden koehenkilöiden monitorointidataa, käytimme lineaarisia regressiomalleja ja kernel-regressiota verensokerin aikasarjan mallintamiseen. Regressiomallien sovittamisen jälkeen erotimme malleista regressiokertoimet, joita käytimme yksilön verensokerin dynamiikan piilevää rakennetta kuvaavina verensokeriprofiilivektoreina.

Erotettu verensokeriprofiili lisäsi tarkkuutta glykeemiseen vaihteluun perustuvassa T2D ja esidiabeetikoiden luokittelussa. Korkean verensokerin ennustamisessa verensokeriprofiili ei tuottanut merkittävää parannusta mallin suorituskykyyn.

Diplomityön tulosten perusteella regressikertoimet voivat mahdollisesti sisältää merkityksellistä tietoa yksilön verensokerin käyttäytymisestä. Saatuja tuloksia tulee tarkastella kriittisesti pienen otoksen (N=62) ja CGM-teknologian mahdollisten epätarkkuuksien vuoksi. Lisäksi saatuja tuloksia tulisi verrata olemassaoleviin lääketieteellisiin glukoosin vaihtelevuuden arviointimenetelmiin.

Asiasanat: jatkuva glukoosinseuranta, verensokeri, diabetes, esidiabetes, ohjattu oppiminen

Kieli: Englanti

Acknowledgements

I am deeply grateful to my thesis advisor Sandor Szedmak for his patient guidance, extensive feedback and support during the thesis. I also want to thank professor Juho Rousu for the comments, help with scoping the thesis topic and supervising my thesis.

Moreover, I want to thank Tiinamaija Tuomi and Liisa Hakaste from University of Helsinki, for sharing their knowledge and expertise and providing access to a clinical dataset collected in Botnia Study.

Thesis the work was performed at the Aalto University, Folkhalsan Research Center and Institute for Molecular Medicine Finland, University of Helsinki. This thesis was funded by Folkhalsan Research Center.

Espoo, May 23, 2021

Viljami Raiski

Abbreviations and Acronyms

CGM	continuous glucose monitor
T2D	type II diabetic
T1D	type I diabetic
RMSE	Root Mean Squared Error
IGT	impaired glucose tolerance
AP	artificial pancreas
DCT	discrete cosine transformation

Contents

Chapter 1

Introduction

Diabetes mellitus is a growing problem in both developing and developed countries, affecting the lives of over 422 million people worldwide [52]. Alone in the United States alone, around nine per cent of the population are diagnosed with diabetes, and over 80 million prediabetics are at high risk of progressing to type 2 diabetes [29]. To a large extent, the diagnosis of blood glucose dysregulation is based on single-time point measurements. However, nowadays, commonly used innovations, such as continuous glucose monitors (CGM), can provide access to time series of blood glucose data. Blood glucose time series data has been utilised in various machine learning applications such as predicting blood glucose values [34, 42], developing new approaches for detecting and assessing glycaemic variability [16, 21] and classifying and diabetes and prediabetes [26, 27, 41, 57]. Most of the current research has been focusing on the applications for subjects diagnosed with T1D due to multiple clinical applications, such as the artificial pancreas. Less focus has been on machine learning related research on blood glucose time series of T2D, prediabetic and healthy subjects. Recent research has been focusing on finding new diagnostical patterns from the CGM data. For example, results in study by Matabuena et al. [32] might indicate that predicting glucose behaviour evolution with continuous monitoring can provide more information than widely used traditional diabetes biomarkers. Furthermore, findings in a study by Hall et al. [16] suggest that the standard way of using single time point measures to diagnose diabetes and prediabetes might miss subjects with a high risk of diabetes and prediabetes compared to the analysis of CGM data. In medical literature, glycemic variability, that is, dynamic properties of blood glucose, has become an established theme. Multiple studies indicate that different characteristics of blood glucose time series have been correlated with an increase in severity and frequency of different unfavourable outcomes, such as microvascular complications and cardiovascular disease [18, 23, 36].

The thesis suggests that the coefficients of a blood glucose forecasting model could capture meaningful information about individuals' blood glucose dynamics. However, further research is needed about the effect of food intake in modelling blood glucose dynamics and the meaningfulness of the information captured by the forecasting model in medical applications.

1.1 Thesis scope

This thesis explores machine learning approaches of modelling blood glucose dynamics and examines whether blood glucose modelling can capture clinically meaningful blood glucose behaviour. The main focus of the thesis is not the prediction of blood glucose values or the classification of the patient's diagnosis based on the blood glucose dynamics but rather explore machine learning methods to capture the latent structure in blood glucose dynamics and test if the latent structure has applications in medical use-cases.

The thesis covers two main topics:

- Train a time series forecasting model for blood glucose and food intake time series data
- Extract regression coefficients from the time series forecasting models and apply them in diabetes/prediabetes classification and hyperglycemia prediction

Thesis scope and workflow is visualised in Figure 1.1.

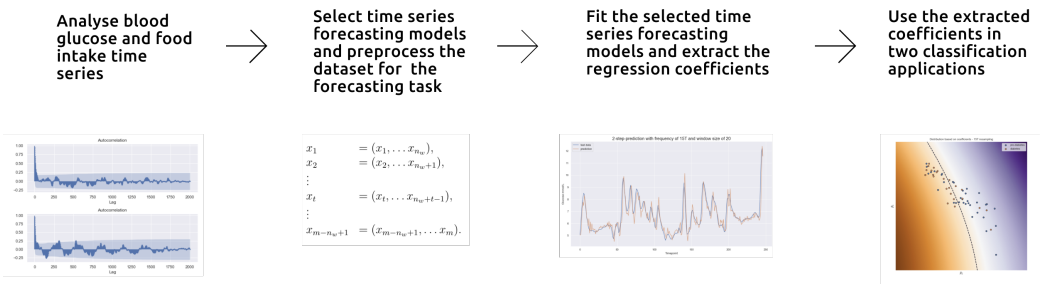

Figure 1.1: An overview of the thesis scope and workflow The first phase was to analyse the dataset's structure. After the analysis, time series forecasting methods were selected and the dataset preprocessed. After selecting the suitable methods, time series forecasting models were trained. In the final part, the regression coefficients were utilised in two medical use-cases.

1.2 Thesis structure

The thesis has the following structure. In Chapter 2, we build a basis for understanding the thesis problem environment. Introduction to blood glucose, blood glucose dysregulation and continuous glucose monitoring is given. After that we go through the relevant literature of current machine learning applications using blood glucose time series data. Chapter 3 describes the theoretical basis of the selected methods and the materials used in the experimental part of the thesis. In addition, the method for extracting latent structure information of blood glucose is described. In Chapter 4 we present the results achieved in blood glucose forecasting task. In Chapter 5 we look into the applications of the produced blood glucose profiles. The blood glucose profiles are used in two use-cases, hyperglycemia prediction and T2D/prediabetes classification. Chapter 6 contains the further analysis of the results, recommendations for future studies and conclusions.

Chapter 2

Background

The utilisation of continuous glucose monitors has enabled more convenient access to blood glucose time series and increased research on machine learning applications using blood glucose time series. Due to the numerous factors affecting blood glucose patterns and different possible use-cases in the medical field, many machine learning models have been utilised in various medical applications. We first go through the biological background for understanding blood glucose dynamics and blood glucose-related illnesses. After that, we describe the function of continuous glucose monitoring systems. In the final section, we look closer at the different machine learning applications using CGM data.

2.1 Blood glucose and glucose dysregulation

In the following sections, concepts of blood glucose, glucose metabolism and different forms of blood glucose dysregulation.

2.1.1 Blood glucose

Blood glucose levels refer to the glucose concentration present in the blood. Glucose concentration is regulated by insulin hormone, which promotes glucose absorption from blood to liver, fat and skeletal muscle cells [46]. The international standard way of measuring blood glucose concentration is molar concentration, defined as mmol/L. In some countries, such as the United States, the concentration is measured as mass concentration mg/dL. Individual blood glucose levels are affected by multiple factors. These factors can be divided into three separate factors. First, there are common factors, including previous blood glucose levels, amount of food intake, insulin intake,

pregnancy, alcohol intake, smoking and drug and vitamin intake. The second group of factors is individual factors, including physical exercise load, dawn phenomena and menstruation. Finally, there are unpredictable factors, including stress and concomitant diseases [5]. American Diabetes Association also mentions dehydration and short- and long-term pain as blood glucose affecting factors [3].

In addition to the factors mentioned above, different metabolic disorders, such as diabetes mellitus, can cause glucose dysregulation [46]. Glucose dysregulation refers to abnormal blood glucose behaviour such as high blood glucose and low blood glucose. also called hyperglycemia and hypoglycemia.

2.1.2 Diabetes and prediabetes

Diabetes mellitus refers to insufficiency of insulin or in its functioning [46].

The most common division of diabetes is dividing it into two types; Type 1 Diabetes and Type 2 Diabetes, although some studies have also proposed a new classification for adult-onset diabetes containing five different types of diabetes [2]. Type 1 diabetes is often called insulin-dependent diabetes, and it usually appears in childhood [46]. In Type 2 diabetes (T2D) body can produce insulin, but due to the loss of insulin receptors in cell membranes, insulin cannot exert its effect on cells.

Another common form of a metabolic disorder is prediabetes, which refers to a state where individuals blood glucose levels are higher than normal level but not yet as high as type 2 diabetes. Elevated blood glucose levels are caused by insulin resistance, which means that cells are not able to absorb glucose properly [8]. The prediabetic state can be separated into two different states, impaired glucose tolerance (IGT) and impaired fasting glucose (IFG). IFT refers to a state where fasting glucose is over 6.1 mmol/l, and IGT refers to a state where two-hour glucose levels are between 7.8–11.0 mmol/l [44]. The risk of progressing to diabetes is quite high for prediabetic patients, and approximately 25 % of prediabetics progress to type 2 diabetes in 3-5 years after the diagnosis [37].

The standard way of diagnosing diabetes and prediabetes has been using single time point measures such as fasting glucose, oral glucose tolerance test or HbA1c, glycated haemoglobin. The weakness of the single-time-point approach is that it might miss subjects with a high risk of diabetes and prediabetes [26]. In medical literature, glycemic variability, which is the dynamic properties of blood glucose, has become an established theme. Different glycemic patterns have been correlated with the increase in severity and frequency of different unfavourable outcomes, such as microvascular complications and cardiovascular disease [18, 23, 36].

2.2 Continuous glucose monitoring

The arrival of continuous glucose monitoring sensors has enabled a novel way of measuring glucose dysregulation and capturing intracity of individuals blood glucose patterns [16]. These sensors measure and store individuals' glucose concentrations every 1-15 minutes for several days, depending on the sensor manufacturer [13]. Continuous glucose monitoring sensors have enabled access to glucose time series data. Compared to the traditional self-monitoring of blood glucose, where glucose concentration is measured from the blood using, CGM sensors measure glucose concentration from interstitial fluid [13]. On a high level, the CGM system consists of two components, a worn sensor that collects blood glucose readings automatically with different time frequencies and a reader that either by scanning or through wireless transmission collects and displays the sensor data. Most of the currently used CGM systems rely on an electrochemical blood glucose biosensor, which senses blood glucose via a single electrode applied under the skin [17]. The electrode is covered with an enzyme that reacts with glucose. The sensor generates current based on the contact between glucose molecules and the enzyme coated electrode, and the current generated is proportional to the amount of blood glucose molecules in contact with the electrode [54]. When the current is converted to a blood glucose concentration, the sensor can transmit the blood glucose data, for example, via Bluetooth[13].

2.2.1 Limitations

One of the CGM technology limitations is the time difference between the readings of the CGM sensor and the actual blood glucose values. The time lag between interstitial glucose values of a CGM and actual blood glucose concentrations is approximately 10–15 min [25].

Compared to the current standard of care method point-of-care capillary glucose testing (POC), the CGM technology can result in different blood glucose readings. In their study, Galindo et al. [14] showed that the mean daily glucose was 12.8 mg/dL (0.7 mmol/L) lower when measured with FreeStyle Libre Pro CGM compared with POC. Although the readings from CGM were significantly lower, the clinical accuracy was still at an acceptable level. Some of the manufacturers offer systems, for example, in the Guardian Real Time or iPro CGM systems, that have a possibility to calibrate the CGM reading by using the actual blood glucose values as calibration values, reducing the error margin of the sensor [33].

2.3 CGM data and machine learning applications

One way to define machine learning is to autonomously learn from the input data and the capability to detect significant patterns from the input data. One common way to categorize machine learning is to divide the applications into supervised and unsupervised learning [48]. Supervised learning can be described as a task of learning a classifier using labelled data. In contrast, in unsupervised learning, there are no target labels we try to predict, and the aim is to organize and summarise the data in a meaningful way.

In addition to the above mentioned, reinforcement learning is proposed as a third paradigm [51]. The reinforcement learning approach is characterized by its focus on goal-directed learning from interaction.

Use cases of machine learning using data provided by CGM data range from detecting new patterns of glucose dysregulation [16], diabetes and prediabetes classification [1, 26, 27], blood glucose anomaly detection [12, 15, 22, 30, 47] blood glucose level prediction [30, 34, 42, 58], insulin bolus calculation [38] to blood glucose control algorithms [9, 10]. Blood glucose has been a widely researched topic in the machine learning research field, but there has been less emphasis on machine learning applications utilising data provided by CGMs. In the following sections, we cover supervised, unsupervised and reinforcement learning applications that have utilised CGM data in their work.

2.3.1 Supervised learning applications

One of the natural use cases of blood glucose time series provided by CGM is time series forecasting. In the domain of blood glucose forecasting, Xie and Wang [58] conducted experiments with various machine learning approaches on the Blood Glucose Level Prediction Challenge (BGLP). The experiments were conducted using the OhioT1DM Dataset. The OhioT1DM dataset contains the following data for each patient (N=6): a blood glucose time series from a CGM with 5-minute intervals, finger sticks blood glucose levels, bolus and basal insulin doses, self-reported times of exercise, work and sleep and 5-minute aggregations of heart rate, step count, galvanic skin response (GSR), skin temperature, and air temperature [31]. In the prediction task, meal sizes, measured heart rate, insulin delivery rate and glucose were used as features. In their experiments, two types of predictions were performed; recursive multi-step and direct multi-step. For recursive multi-step prediction, the best performing models were Linear Regression (ARX) and Ridge

regression. When experimenting with direct multi-step prediction, the Support Vector Regression with Radial Basis Kernel had the best performance. The results were compared in terms of Root Mean Squared Error (RMSE). In their study, Bunescu at al. [7] showed that using the physiological blood glucose modelling could be utilised in generating features for blood glucose level predictions. Plis et al. [42] investigated further the applicability of the findings of Bunescu et al. [7] in blood glucose predictions. Using CGM data collected from 5 T1D patients, three forecasting models were produced, ARIMA, SVR and SVR with ARIMA features. The ARIMA model was trained with four days of CGM data. The SVR models were trained by using physiological states in different time points and blood glucose trends, i.e. the difference between blood glucose values between time points. In addition, the SVR model that utilised features extracted from the ARIMA was used for the prediction, which resulted in the best results in blood glucose predictions for the 30-minute and 60-minute horizon. CGM data has also been utilised in artificial neural network approaches in blood glucose level prediction. In their study, by Pappada et al. [40] trained neural network to predict the entire vector of blood glucose values 75 minutes ahead using 5-minute interval CGM data and data from an electronic diary including nutritional intake, insulin dosages, lifestyle/activities, emotional factors and symptoms caused by hypoglycemia or hyperglycemia. Another blood glucose prediction approach proposed by Mhaskar et al. [34] utilised only the five-minute interval of blood glucose data for training deep learning model on two tasks, predicting blood glucose value 30 minute ahead and predicting the rate of blood glucose change after 30 minutes from now.

In the domain of blood glucose anomaly detection, Jensen et al. [22] trained a support vector machine (SVM) to detect hypoglycemia in real-time events from professional CGM data. The dataset used contained CGM data of 10 male subjects. As a result, the best model detected 100% of the hypoglycemic events with one false positive. Furthermore, Georga et al. [15] utilised SVR on predicting nocturnal and diurnal hypoglycaemic events. The dataset contained glucose data of 15 patients with type 1 diabetes under free-living conditions. In their experiments, two prediction horizons, 30-minute and 60-minute horizons were used. The resulting model was able to detect nocturnal hypoglycemic events with a sensitivity of 94% on both horizons. For the diurnal hypoglycemia prediction, the prediction sensitivity was 92 % for the 30-minute horizon and 96 % for the 60-minute horizon.Hypoglycemia prediction was also studied by Plis et al. [42], and in their experiments, the ARIMA-model and SVR capabilities on hypoglycemia prediction were compared with diabetes expert predictions. The dataset used contained 5-minute interval CGM data for five T1D subjects and corresponding data on

each subjects insulin intake, meals, exercise and sleep. The ARIMA model was trained using CGM data only, and the SVR was trained using CGM data and physiological model-based features, such as meal absorption dynamics, insulin dynamics and blood glucose dynamics. The resulting SVR extended with physiological features outperformed diabetes experts in the hypoglycemia prediction task and correctly predicted 23% of the hypoglycemic events. Although the methods above have not been used in real-life applications, some applications have been utilised in clinical use-cases. For example, in their patent, Roche Diagnostics Operations Inc. [11] described using Gaussian Process regression as a method for predicting future hypoglycemia events from CGM data of a diabetic person. Overall, hypoglycemia classification seems to be a more investigated topic compared to the hyperglycemia prediction. This conclusion is supported by a literature review conducted by Woldaregay et al. [57] on machine learning applications in type 1 diabetic use-cases. From the 47 studies included in the review, five were related to hyperglycemia prediction, 39 were related to hypoglycemia prediction and three related to the glycemic variability classification.

CGM data has also been utilised in the diagnosis of diabetes mellitus using supervised machine learning approaches. In their study, Acciaroli et al. [1] trained a machine learning model for healthy, prediabetic and diabetic patient classification using glycaemic variability indices. Glycaemic variability (GV) indices characterise the variation in blood glucose profiles extracted from the data provided by a CGM sensor. For every 102 subjects in the dataset, 25 CV indices were extracted. After feature selection, they chose eight GV indices for healthy/IGT&T2D classification and five GV indices for IGT/T2D classification. Using the chosen subset of GV indices as input for logistic regression, they achieved 91.4 % accuracy in healthy vs IGT&T2D patients and 79.5 % accuracy in IGT vs T2D classification. The reported accuracies were mean of five-fold cross-validation. The classification between prediabetic and T2D was investigated further by Longato et al. [26], where a support vector machine with polynomial kernel was developed to detect subjects affected by IGT and T2D. The best T2D/IGT classification model was SVM with polynomial kernel trained with 37 glycemic variability indices and four additional parameters: age, sex, BMI, and waist circumference. The achieved four-fold cross-validated accuracy achieved was 72.5 %. The research of T2D/IGT classification was continued in the study by Longato et al. [27], where resulting four-fold cross-validated accuracy of the best model was 69.7 %. The best model was an SVM using linear kernel trained using 17 glycemic variability indices, age, sex, BMI, and waist circumference.

2.3.2 Unsupervised learning applications

CGMs have enabled closer inspection of the patterns in blood glucose variability and new ways of characterising and comparing blood glucose between individuals. In their study, Hall et al. [16] analysed CGM traces of 57 subjects who wore the CGM for a minimum of two to four weeks. The dataset was preprocessed into two and half hour time windows and then using the resulting set of 2,5 hour time windows clustered using spectral clustering. After analysing the clustering results, three different clusters were found, suggesting three different types of glycemic variability. The three clusters, low variability, moderate variability and severe variability, were then used to assign a glucotype for each subject, describing which of the three distinct blood glucose patterns occurred most of the time in their CGM traces.

A similar approach was used by Inayama et al. [21], where hierarchical clustering was used to assess differences in blood glucose patterns of healthy women and women with gestational diabetes (GDM). After clustering two-week-long CGM trace collected from 29 pregnant women, researchers found three distinctive clusters, low glucose variability, moderate-to-high glucose variability, and high glucose variability. Further analysis revealed that most of the subjects diagnosed with GDM were in the group of high glucose variability, indicating that clustering could help to characterise blood glucose profiles between women with GDM and healthy women.

2.3.3 Reinforcement learning applications

One recently researched application of reinforcement learning (RL) based algorithms is blood glucose control algorithms. Blood glucose control algorithms are one primary component of the artificial pancreas (AP). AP is a safe and effective approach for treating diabetes mellitus and consists of three main components [53]:

1. A CGM measuring blood glucose levels continuously
2. An insulin pump delivering the correct amount of insulin needed
3. Blood glucose control algorithm that uses the measured blood glucose levels from CGM to instruct the insulin pump on insulin delivery

RL algorithms have been shown to provide an intelligent, tailored and precise way of computing insulin delivery [53]. However, despite the wide usage of RL algorithms in AP applications, only a fraction of the studies have utilised clinical CGM data in their studies. In their review, Tejedor

et al. [53] found that only one study over all the studies selected for review (N=31) utilised clinical CGM data. Furthermore, most of the current studies are performed in silico and thus not validated for clinical use. The amount of research utilising RL approaches in blood glucose regulation problems has increased recently. However, in the current literature, there is not that much focus on utilising additional factors that affect blood glucose, such as physical exercise and meal intake [53].

2.3.4 Summary

Most machine learning research utilising CGM data focuses on data gathered from subjects with type 1 diabetes, leaving only a tiny amount of research on other types of diabetes and blood glucose dysregulation such as T2D, GDM, IGT, and IFT. Supervised learning methods have been widely used in various use cases. Based on the current literature, there is no machine learning approach working for every problem, and the selected machine learning approach should be problem-specific. In addition, many of the studies had a low sample size, which reduces the generalisation of the achieved results. The focus of the unsupervised methods has been on detecting interesting glycemic patterns from blood glucose time series. Among the Reinforcement Learning approaches, the most focus has been on glucose control algorithms, which are essential for developing the artificial pancreas for people diagnosed with T1D.

In reinforcement learning applications, most studies have focused on using blood glucose observation as a predictor. However, many of the supervised learning approaches applied in offline learning setup have incorporated additional data, such as physiological models or meal intake data, in their models.

The existing literature has shown a correlation between frequency and severity of medical complications and dynamics properties of blood glucose [18, 23, 36] and experiments by Acciaroli et al. [1], Longato et al. [26] and Longato et al. [27] have shown the usefulness of glycemic variability indices in classifying subjects with T2D and prediabetes. Based on these findings, we hypothesise that the information captured by blood glucose predictive models could provide insights into an individual's blood glucose dynamics, thus help to assess, for example, individuals physiological state such as T2D or prediabetes. Using the information captured by blood glucose forecasting models is analogous to the experiments of Inayama et al. [21] and Hall et al. [16] for blood glucose characterisation. However, instead of unsupervised learning methods, in this thesis, we uses supervised learning methods to learn the latent structure of blood glucose time series.

Chapter 3

Methods and Materials

In this chapter, we introduce and describe the theoretical basis of the methods used for extracting the blood glucose characterisation and methods used to assess the usefulness of the resulting blood glucose characterisation in classification tasks (Sections 3.1 and 3.2). After that, description of the dataset and analysis of the dataset is provided in sections 3.3, 3.4. Last sections 3.5, 3.6 and 3.7, describe the data preprocessing before fitting a predictive model, extraction of blood glucose profile and overview of the blood glucose extraction algorithm.

3.1 Time series forecasting for blood glucose forecasting

The first task in extracting different blood glucose patterns in this study was fitting a time series forecasting model from which we can extract the information the predictive model has captured. For this purpose, basic supervised machine learning approaches, Linear Regression and Kernel Regression were applied to the forecasting problem.

3.1.1 Linear Regression models

Linear regression can be described as a task of predicting real valued labels based on training data, and it assumes linear relationship between the labels and training data. Given real valued labels $\{y_1, y_2 \dots y_n\}$ and input vectors $\{x_1, x_2 \dots x_n\}$ with p-features, the Linear Regression for each $i = 1, 2 \dots n$, is formulated as:

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + ... + \beta_p x_{i_p} + \epsilon_i \tag{3.1}$$

Equations for each $i = 1, 2 \ldots n$ can also be expressed in matrix notation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

where $\boldsymbol{y}$ corresponds the target real valued labels, $\boldsymbol{X}$ is matrix of input variables containing bias and a p length feature vector, $\boldsymbol{\beta}$ is a regression coefficient vector of $p + 1$ elements, and $\boldsymbol{\epsilon}$ is a vector for noise terms.

For the Linear Regression task, the most commonly used loss function defining the difference between the predicted value and ground truth is the squared loss, also known as L_2 loss. Squared loss is defined as:

$$L_2(\hat{y}, y) = |\hat{y} - y|^2 \tag{3.3}$$

The empirical risk minimization related to the linear regression is the mean squared error (MSE). The MSE is defined as:

$$\text{MSE}(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^{n} L_2(\hat{y}_i, y_i) \tag{3.4}$$

Using the Ordinary least squares estimator, the coefficient parameter $\boldsymbol{\beta}$ can be estimated with a closed-form solution:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{3.5}$$

The resulting coefficient vector $\boldsymbol{\beta}$ is then considered as a blood glucose characterization vector.

The basic setup in blood glucose forecasting is using the past glucose values to predict the future values. The Linear Regression model using only the past values is identical to an autoregressive model [20]. Multiple studies, such as, Plis et al., Bunescu et al., Saiti et al. and Xie and Wang [7, 42, 45, 58] have demonstrated the applicability of autoregressive models in blood glucose forecasting by numerous studies. We justified the selection of the Linear Regression model based on the similarity with the autoregressive model and the support demonstrated by earlier research on the applicability of an autoregressive model in similar blood glucose forecasting tasks.

3.1.2 Kernel Regression

Kernel Regression has been utilised in blood glucose forecasting task in multiple stidues. For example, Plis et al., Bunescu et al., and Wiley et al [7, 42, 56] applied Kernel regression approaches in their blood glucose forecasting experiments. The main benefit from the Kernel methods is that it allows applying

linear learning methods to nonlinear problems [48]. This is achieved using a method called the kernel trick.

In kernel trick, the input features are implicitly mapped into a higher dimensional space [48]. This allows implicit computation of inner product between the input features, which would otherwise be computationally expensive. This kernel trick is achieved via kernel functions. Given domain space X and an embedding ϕ of all x and x' in X into a Hilbert space, kernel function is:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle \tag{3.6}$$

where $\langle ., . \rangle$ is an inner product. Often used kernel functions contain linear kernel, polynomial kernel and Gaussian kernel. In this thesis, linear kernel and polynomial kernel functions were used.

In this thesis, we use the Kernel Ridge Regression implementation provided by scikit-learn.In Kernel Ridge Regression, the goal is to combine Ridge regression with a kernel trick [55]. Using the mean squared error, the Kernel Ridge Regression optimization problem can be formulated as:

$$\underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{m} (\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y})^T (\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha} \tag{3.7}$$

where $\boldsymbol{K}$ represents the kernel of matrix input data $\boldsymbol{X}$, $\boldsymbol{y}$ is the real valued target label, $\boldsymbol{\alpha}$ is the weight vector. The first terms corresponds the error norm and λ is a regularisation term.

The dual optimization problem has an analytical solution defined as

$$\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y} \tag{3.8}$$

The resulting $\boldsymbol{\alpha}$ is then considered as blood glucose profile extracted from the Kernel Regression model.

3.2 Diabetes and prediabetes classification and hyperglycemia prediction

After extracting the pattern encapsulating information from the time series forecasting model, this information's usefulness needs to be tested in medical use-cases. For this purpose, two applications were selected. The first use-case was to use the extracted information as a feature when classifying subject prediabetic or T2D. This use-case was similar to the studies by Acciaroli et al. [1], Longato et al. [26] and Longato et al. [27], where glycemic variability indices were used for the T2D/prediabetes classification task. The second

approach is to use the extracted glucose profile for hyperglycemia prediction. Both of the use-cases are classification tasks, and therefore Logistic Regression was used in both applications was Logistic Regression.

Logistic Regression is one of the most widely used machine learning method for classification tasks. Since one of the extracted blood glucose profile applications is the classification of T2D and prediabetic patients, we use the same method used by Acciaroli et al. [1] in a similar task.

Logistic Regression belongs to the family of linear predictors, and it can be used to model probabilities of output classes, given the input features. In Logistic Regression, probability of a output class is modeled using a sigmoid function called *logistic function* defined as [48]:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.9}$$

In two class classification $y \in \{C_1, C_2\}$, given input vector x, the probability of each class can be formulated as:

$$\begin{cases} P(y = C_1|x) = \frac{1}{1+e^{-w^T x}} \\ P(y = C_2|x) = 1 - P(y = C_1|x) \end{cases} \tag{3.10}$$

Given target labels $\{y_1, y_2 \dots y_n\}$ and input vectors $\{x_1, x_2 \dots x_n\}$, the empirical risk minimization problem for logistic regression is defined as:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} \log(\exp(-y_i(x_i^T w)) + 1) \tag{3.11}$$

which corresponds the maximum likelihood estimation of the weight vector w.

In this thesis, we use Logistic Regression with l_2 regularisation provided by scikit-learn, which then leads to minimization of the following cost function [6]:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(x_i^T w + c)) + 1) \tag{3.12}$$

To optimize the Equation 3.12 we used coordinate descent algorithm based liblinear-method implemented in scikit-learn.

3.3 Dataset description

The used dataset contains 62 subjects, 31 male, 31 female and aged from 44 to 75. From the total 62 subjects in the dataset, 36 were diagnosed with

IGT, that is, prediabetes and 26 were diagnosed with T2D. The subjects diagnosed with T2D have had their diagnosis for a relatively short period of time, a maximum of three years. The data was collected from subjects who participated in the Botnia Prospective Study and the Botnia PPP Study approved by the ethics committee of Helsinki University Hospital. The subjects were monitored during the EU FP7 Mosaic Project. The dataset contains data collected in two visits: a baseline visit in 2014 and a follow-up visit in 2015. During each visit following data was collected:

- Continuous glucose monitoring (CGM) data for 7 days. Each subject was monitored by either Guardian Real Time or the iPro CGM systems. Format of the CGM data was time series with 5 minute intervals (See Figures A.1 A.2 in Appendix A).
- Food diary containing timestamps of food intake and approximation of nutrients.
- Exercise diary containing timestamps of exercise start time and a description of the exercise.
- Additional form containing the following data: visit date, age, height, sex, diagnosis, blood glucose and insulin collected during OGGT and blood glucose during test meals.

Height, weight, age and BMI distributions are summarised in Figure 3.1. Noteworthy is that a significant amount of the subjects are overweight given the fact that threshold value of BMI for overweight is 25 [35].

3.4 Glucose time series analysis

The CGM data in the dataset was collected from two visits, baseline and follow-up. For both visits, approximately seven days of blood glucose data were collected, resulting in two seven day time series of blood glucose.

Mean and standard deviation of blood glucose values by weekdays are summarised in Figure 3.2. For mean glucose no trend is observable in terms of weekdays, and the values for each weekday seem to correspond to each other in both groups, prediabetics and diabetic. The standard deviation seems to remain stable across weekdays among the prediabetic subjects, but among diabetic subjects, the variation between days is higher.

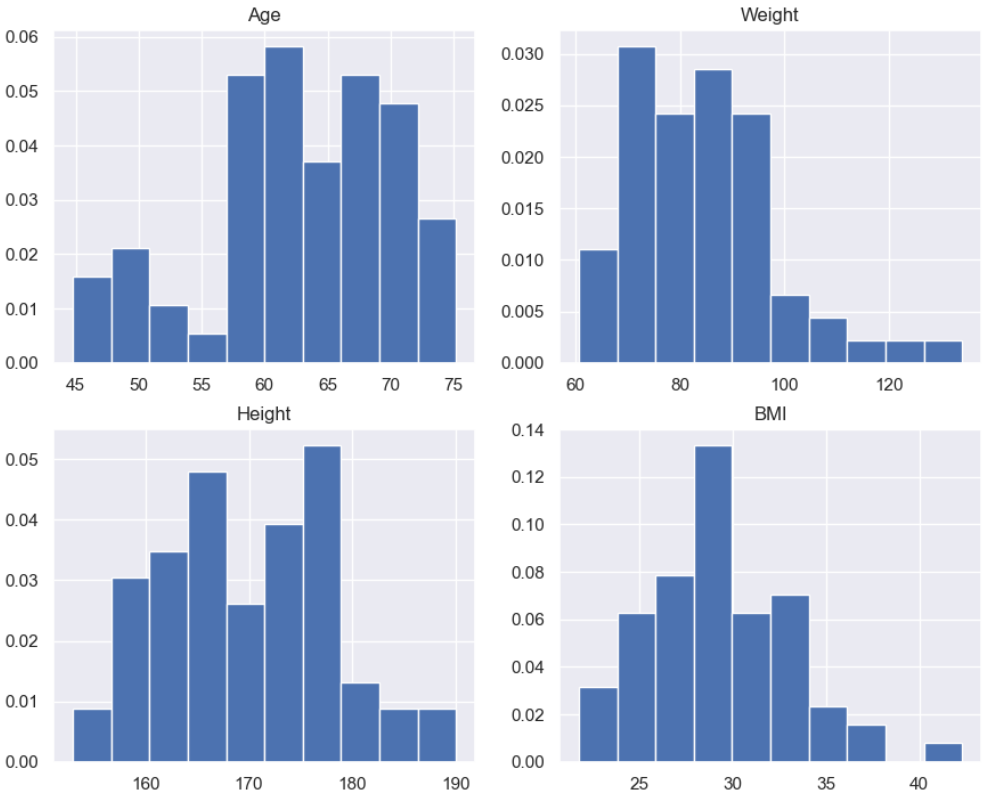

Figure 3.1: Histograms of age, weight, height and BMI of the subjects in the dataset

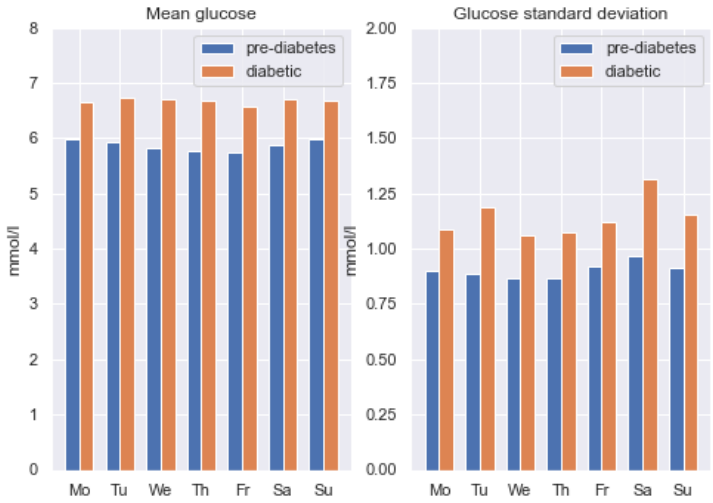

Figure 3.2: Mean blood glucose and blood glucose standard deviation by weekday

3.4.1 Autocorrelation

The topic of interest in the blood glucose time series is to detect underlying patterns and trends. One way to assess the patterns is by exploring autocorrelation within the time series. Autocorrelation is defined as a correlation between the underlying time series signal with a lagged, delayed copy of the same time series [20].

Plotting the autocorrelation revealed a strong correlation between subsequent values and cyclical patterns visible with larger lags. Autocorrelations plots for three subjects extracted from the CGM traces are visualised in Figure 3.3. High autocorrelation is visible near lag of one, and then it decreases fast after it starts to follow an observable pattern of oscillating between positive and negative autocorrelation. Computing the autocorrelation did not reveal any major differences in blood glucose time series patterns between diabetic and prediabetic subjects.

3.4.2 Discrete cosine transformation

In addition to the autocorrelation plots, discrete cosine transformation (DCT) is used to analyse underlying patterns in blood glucose time series provide by CGMs. Discrete cosine transformation is a method to express finite sequence, for example, blood glucose time series, in terms of a sum of cosine

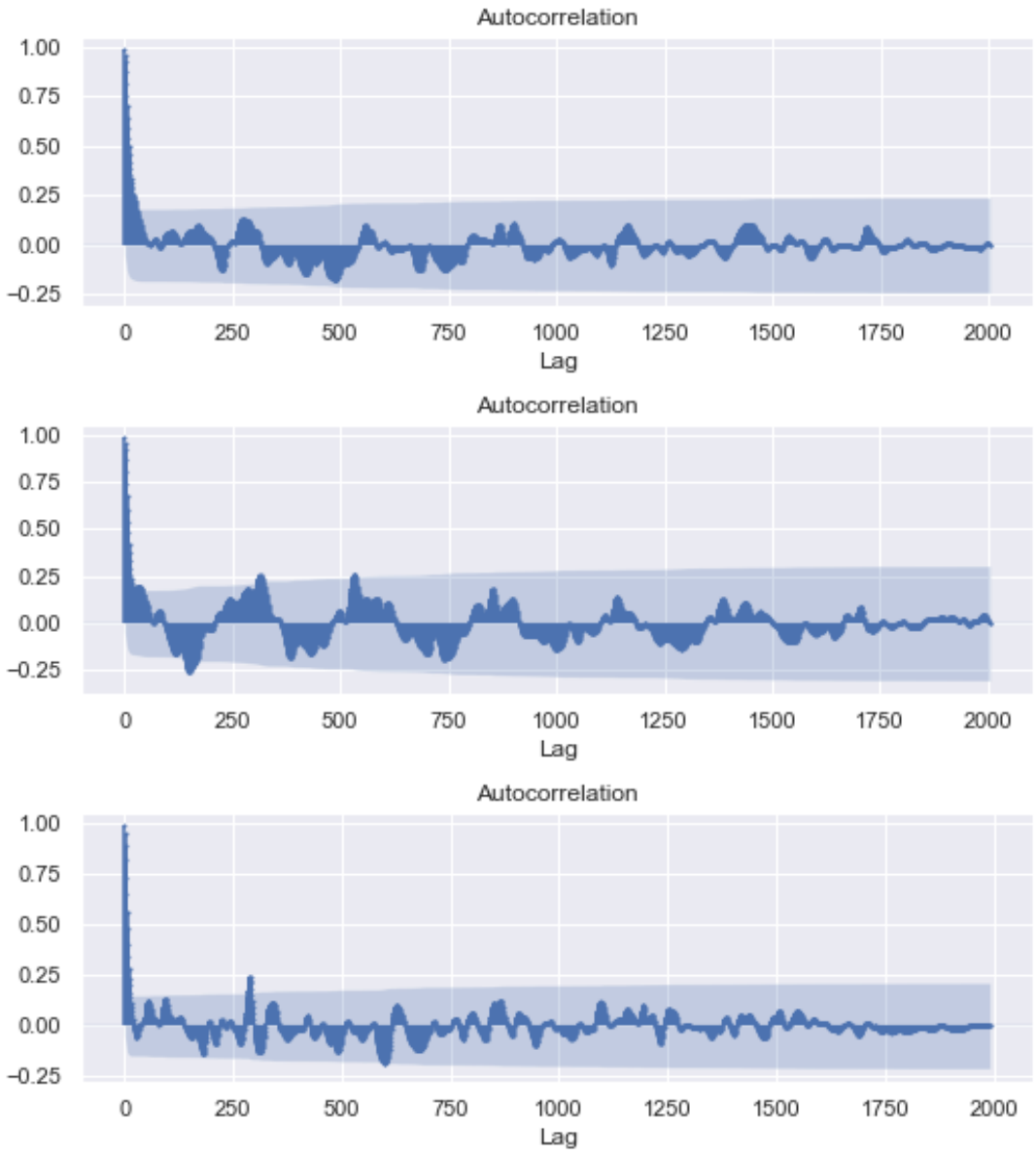

Figure 3.3: Autocorrelation plots with different lags of CGM traces of three subjects. The first and last row plots are extracted from a subject diagnosed with T2D, and the middle one is extracted from a subject diagnosed with prediabetes. Run with 5-minute data for seven days of data.

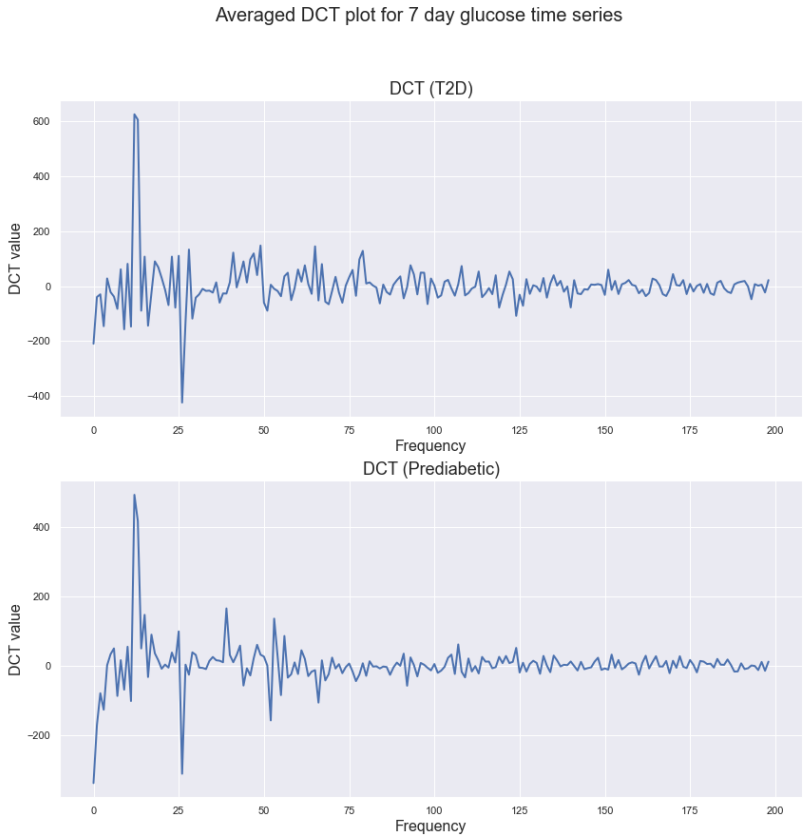

Figure 3.4: Average discrete cosine transformation (DCT). The DCT values plotted are averages over all subjects in the diagnosis group. Ran with 5-minute data. Figure description in Section 3.4.2.

functions oscillating at different frequencies [43]. Results of averaged DCT values between all subjects is visualised in Figure 3.4. Both of the subject groups, diabetic and prediabetic, had high frequency on points 12 and 13, which in 5-minute data represents 1h and 1h 5 minutes. In addition, more high frequencies appear between 20-66, which in 5-minute data corresponds to a range of 1 h 50 min to 5h 30 min.

3.4.3 Comparing glycemic variability indices between diabetic and prediabetic subjects

As described in Chapter 2, one way to assess individuals blood glucose behaviour is to use glycemic variability indices.

We analysed eight glycemic variability indices: mean glucose, standard deviation, coefficient of variation, time in range, time above range, time below range, LBGI and HBGI. Acciaroli et al. [1] used these indices in their experiments in T2D/IGT classification. Mean glucose is the mean value of all glucose readings, coefficients of variation is defined standard deviation divided by the mean glucose, time in range refers to the percentage of time subjects' blood glucose values are between 3.9 mmol/L and 10 mmol/L [1]. The time below range refers to the percentage of time the blood glucose has been under 3.9. mmol/L and time above range refers to the percentage of time having blood glucose over 10 mmol/L. HBGI and LBGI refer to high blood glucose index and low blood glucose, respectively.

Differences between prediabetics and diabetics in eight commonly used glycemic variability indices are illustrated in Figure 3.5. The difference is notable in two indices, "Time above 10 mmol/L" and "Time below 3.9 mmol/L". Those indices represent time spent in hyperglycemia and hypoglycemia, respectively. Based on the observations from Figure 3.5, we can see differences in the blood glucose dynamics between prediabetic and diabetic subjects. This thesis aims to capture similar dynamics with the models presented in Section 3.1.

3.4.4 Glucose time series and events

In addition to the blood glucose time series, the dataset contains a food diary where subjects have written down their meals. In addition to the food intake time, the diary contains estimations of the amount of nutrients for each meal. As described in Chapter 2, food intake is a factor with the highest impact affecting blood glucose behaviour. Therefore it should be considered when forming a model for predicting blood glucose.

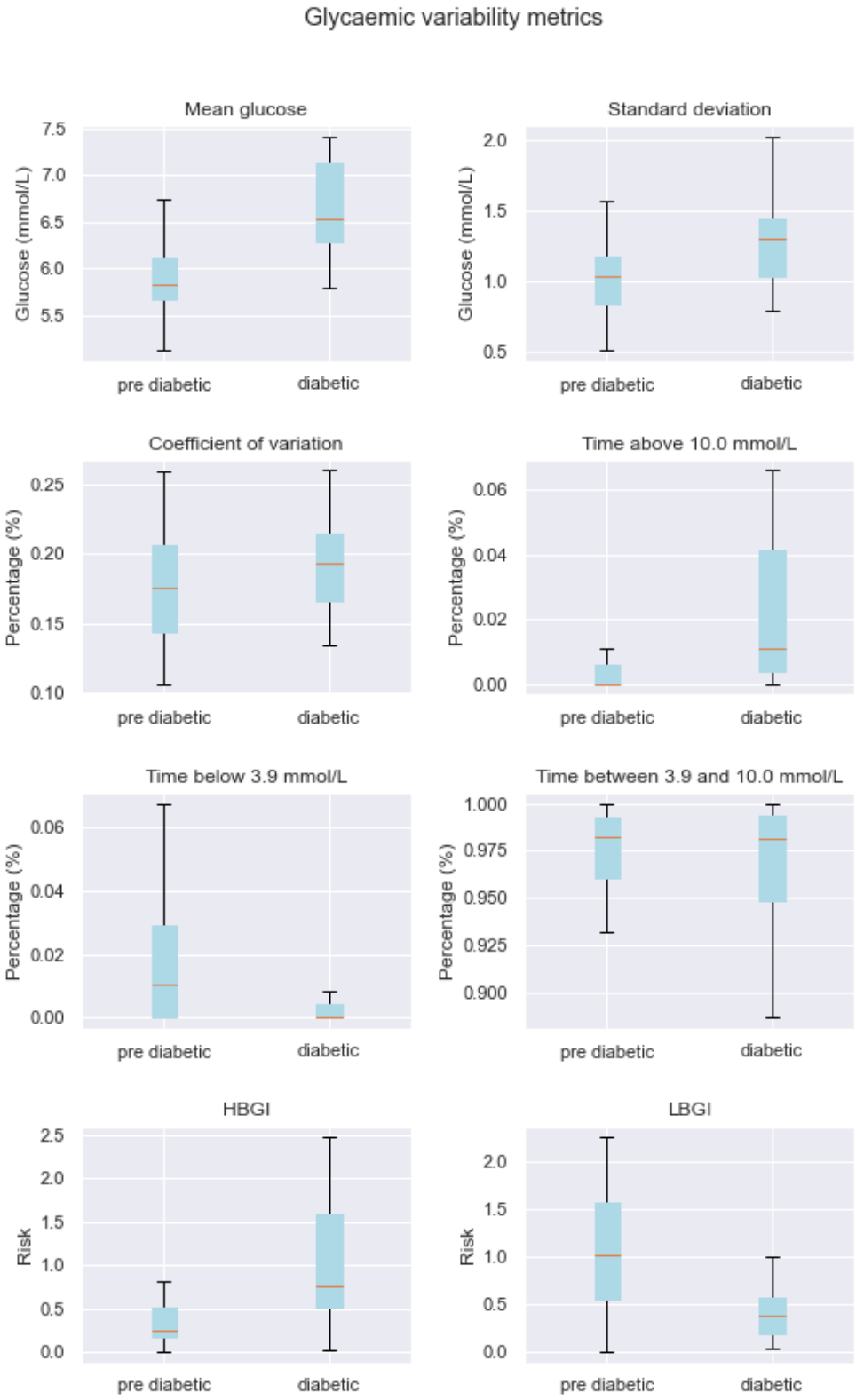

Figure 3.5: Summary of glycemic variability metrics computed for diabetics and prediabetic subjects.

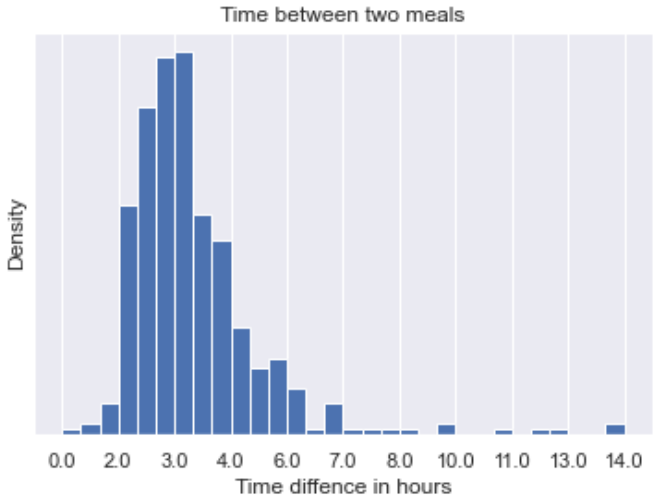

Figure 3.6: Time difference between two subsequent meals

To gather information on the frequency of meal intake impact, the average time between two meals was investigated and compared to the results of DCT plots. The average time between meals is illustrated in Figure 3.6. Most of the data lie between two and six hours, which approximately corresponds to the highest peaks found by utilising DCT discussed in subsection 3.4.2 and visualised in Figure 3.4.

3.5 Data preprocessing

In order to be able to train the methods described in 3, the data needs to be preprocessed. Preprocessing consisted of interpolating missing data, converting all glucose values into the same units, combining event data with glucose time series, downsampling data and finally forming time windows and target vectors the training datasets. The used time windows, resampling frequencies and the number of steps predicted forward are summarised in Table 3.1.

window size (w)	1,2,3,6,10,15,20,25,30
resample frequency (r-freq)	5 min, 15 min, 30 min, 1 h
n-step prediction (n-step)	1,2,3,4,5,6

Table 3.1: Summary of used data preprocessing hyperparameters

3.5.1 Handling missing data

In the glucose time series dataset, approximately 5% of the data was missing. The missing data was augmented using *interpolate* function provided by *pandas* library. Due to the time-dependant nature of blood glucose data for the interpolation was performed by selecting the nearest non-null data point.

3.5.2 Handling data in different units

For one subject part of the glucose values, under 1% of the data, were in different units, causing abnormal results when computing summary statistics. Most of the glucose values are stored in mmol/L in the dataset, but some of the values were in mg/dl. One mmol/L of glucose corresponds ≈ 18 mg/dL. Those values in the wrong format were converted to mmol/L by dividing the mg/dL value by 18.

3.5.3 Combining event information with glucose values

Given glucose time series G_x with time points might differ from $\tau_1, \ldots, \tau_t, \ldots$, there could be some other time related events. These include for example the time points of the food intake, the start time and the end time of exercises. Let the event happen at the time points $\zeta_1, \ldots, \zeta_j, \ldots$ Those time points might differ from $\tau_1, \ldots, \tau_t, \ldots$, but we can assign them to that point in $\tau_1, \ldots, \tau_t, \ldots$ which happens immediately after those events. Namely, if the event happened in time ζ_j then it is assigned to time τ_t such that $\tau_{t-1} < \zeta_j \leq \tau_t$. In this way we can define the events as time series $(e_1, \ldots, e_t, \ldots)$ with values $\{0, 1\}$ on the same time points:

$$s_j \begin{cases} 1 & \zeta_j \in (\tau_{t-1}, \tau_t], \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

We might have several series of measurements, and events, they can be combined in a similar way outlined above.

3.5.4 Data downsampling

In addition to the original 5-minute glucose time series, the dataset was downsampled to less frequent time series using *pandas*-library resample function for the experimentation phase needs. Downsampling was performed in two steps. First, the glucose time series was split into time windows with different frequencies, 15 minute, 30 minute and 1 hour. The final downsampled

glucose time series was generated by taking the mean of glucose values inside the time windows. Thus in addition to the 5-minute interval of glucose values, 15 minute, 30 minute and 1 hour mean glucose time series were inspected. By downsampling, we trained the blood glucose forecasting models to capture short-term dynamics and long-term dynamics.

3.5.5 Forming time windows and target vector

In order to utilize combined time series data in the forecasting model, the data was first preprocessed in following way:

We are given a time series $S_x = x_1, \ldots, x_t, \ldots x_m$ measured in the time points $\tau_1, \ldots, \tau_t, \ldots, \tau_m$. If the elements of S_x are vectors, e.g. $x_t \in \mathbb{R}^{n_w}$, then we can directly form a data matrix $X \in \mathbb{R}^{n_t \times n_w}$ whose rows are equal to the elements, thus $X_t = x_t$ for all t.

If the elements of S_t are scalar values then we can apply a time-window with length $n_w < m$ to form the data matrix. In this case

Step 1 Form a input data matrix $\mathbf{X} \in \mathbb{R}^{n_t \times n_w}$.

Rows of the matrix $\mathbf{X}$ are given by

$$\begin{aligned} x_1 &= (x_1, \ldots x_{n_w}), \\ x_2 &= (x_2, \ldots x_{n_w+1}), \\ &\vdots \\ x_t &= (x_t, \ldots x_{n_w+t-1}), \\ &\vdots \\ x_{m-n_w+1} &= (x_{m-n_w+1}, \ldots x_m). \end{aligned} \tag{3.14}$$

where, each element x_t is a vector containing glucose value and event data related to the time window of length w. This windowing is reducing the full length, but we assume that the window length is much less then the length of the time series.

Step 2 Form a target vector $\mathbf{y} \in \mathbb{R}^{n_t - n_w + 1}$ with respect to the number of steps we want to predict forward n_{step}.

Namely for each row in $\mathbf{X}$ the corresponding target y_i is defined as.

$$\begin{aligned} &(x_1, \ldots x_{n_w}), y_1 = x_{n_{w+1}} \\ &(x_2, \ldots x_{n_w+1}), y_2 = x_{n_{w+2}} \\ &\vdots \\ &(x_t, \ldots x_{n_w+t-1}), y_t = x_{n_w+t} \end{aligned} \tag{3.15}$$

3.6 Blood glucose profile extraction

In the following section, we describe the implementation of blood glucose profile extraction algorithm. The extraction of blood glucose profile can be thought of as analogous to the method used in [7, 42], where features extracted from the ARIMA model were used to improve the blood glucose forecasting model.

For each subject in the dataset, a blood glucose profile was extracted by fitting a Linear Regression model and Kernel Regression model for a blood glucose prediction task. Both regression models were fitted using different combinations of training data based on different preprocessing hyperparameters, in this case *window size*, *number of steps predicted forward* and *resampling frequency*, discussed in Section 3.5. The Linear Regression model was fitted using a few variations of the dataset, with and without food intake information, log-transformed glucose data, and transforming the target value using a Bernoulli Link function. The Kernel Regression was fitted using fixed-size kernel and with linear kernel and polynomial kernel of degree three.

After fitting the model for each subject, we extract the regression coefficients vector ($\boldsymbol{\beta}$ in Equation 3.5 and $\boldsymbol{\alpha}$ in Equation 3.8), that is the weights of the fitted model. The extracted weight works as a characterization vector of subjects blood glucose patterns.

One problem faced during the experiments was that the blood glucose profile extracted from Kernel Regression, that is, the $\boldsymbol{\alpha}$ in Equation 3.8, increases proportionally to the number of training samples, thus leads to a variable size glucose profile, given different amount of data per subject. This problem was solved by setting the kernel size fixed during the training. Fixed-size kernel training was performed using the following procedure:

Step 1 Select the size of the kernel, say N, and let the number of all time points be T.

Step 2 Warm-up period: take the first N time points, $t_1, ..., t_N$ and solve the kernel ridge regression problem, obtain the dual coefficient vector

Step 3 For time point t in $t_{N+1} \dots t_T$

1. Include the data vector from time t.
2. Delete the data vector belonging to that component of the dual coefficients that has the largest absolute value. That means, on that input, we have the highest error.
3. Solve the ridge regression problem again, and obtain the dual coefficient vector

3.7 Algorithm overview

High-level workflow description for extracting blood glucose profile:

Input Blood glucose time series and event time series

Step 1 Select data preprocessing hyperparameters window size w, resampling frequency *r-freq* and number of steps predicted forward *n-step*.

Step 2 Preprocess the blood glucose time series and event time series as described in Section 3.5 based on selection of Step 2. Obtain the training data $\mathbf{X}$ and the target vector $\mathbf{y}$

Step 3 Fit the predictive model $\mathbf{M}$ using the preprocessed training data $\mathbf{X}$ and target vector $\mathbf{y}$

Step 4 Extract the coefficients from the predictive model $\mathbf{M}$

Output Blood glucose characterization vector $\mathbf{x}_{\text{profile}}$

Chapter 4

Blood Glucose Forecasting

This chapter describes fitting the Linear Regression model and Kernel Regression model for the blood glucose time series forecasting task. In the first section, the performance evaluation metrics are described, and in the second section, the results of blood glucose prediction models are described more in-depth.

4.1 Performance evaluation metrics

The comparison between performance of time series models was done based on *Root-mean-squared error*, *Pearson correlation*, and *Spearman correlation*.

4.1.1 Root-mean-squared error

One way to measure how well the resulting model can predict glucose values is a root-mean-squared error. Root-mean-square error, also abbreviated as RMSE, is a standard way of measuring the error between estimator predictions and actual values [20]. Formally RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \tag{4.1}$$

RMSE is a scale-dependent error metrics. The errors are on the same scale as the dataset and cannot be used to compare results performed on datasets with different units [20].

RMSE has been widely used to compare the performance of different models in blood glucose prediction, for example, by, Plis et al. [42] and Bunescu et al. [7] and Xie and Wang [58].

4.1.2 Pearson correlation

In addition to the prediction error, we want to know whether the prediction follows similar patterns as the actual blood glucose. One way to assess the similarities in terms of patterns is to use the Pearson correlation coefficient. Pearson correlation is used to measure the linear correlation between two variables. Given two datasets X and Y, the Pearson correlation $\rho_{X,Y}$ is formulated as:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{4.2}$$

where, cov is the covariance and σ_X and σ_Y are standard deviations of data sets X and Y respectively [19].

4.1.3 Spearman's rank-order correlation

Spearman's rank-order correlation is a non-parametric formulation of rank correlation. One advantage over Pearson correlation is that the Spearman correlation is more robust to the outliers. The Spearman correlation between two variables corresponds to the Pearson correlation between the ranks of two variables.

For sample size n, the raw data X_i and Y_i are converted to ranks rg_{X_i} rg_{X_i}, the coefficient is computed as:

If all sample n ranks are distinct integers, the data X_i and Y_i are converted to ranks rg_{X_i} rg_{X_i}, the coefficient is computed as [50]:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)} \tag{4.3}$$

where

$$d_i = rg_{X_i} - rg_{Y_i} \tag{4.4}$$

is the difference between two ranks of observation i, and n is the number of observations.

If all ranks are not distinct integers, the formulation is:

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}, \sigma_{rg_Y}} \tag{4.5}$$

4.2 Results

Using the data preprocessing hyperparameters defined in Table 3.1 the first step was to experiment how well the Linear Regression and Kernel Regression

perform in modelling the blood glucose behaviour. Using different resampling frequencies, different n-step predictions and window sizes, we aim to capture different types of blood glucose behaviour. These type range from short term patterns, e.g. using five-minute interval data to predict the glucose value after 15 minutes and long-term patterns, e.g. forecasting the mean glucose after six hours using the past hourly mean glucose values. Kernel Regression was run using polynomial kernel and linear kernel. Total number of combinations different models trained based on data preprocessing hyperparameters in Table 3.1 was 3(models) $\times$ 9(window size) $\times$ 4(resampling frequency) $\times$ 6(number of steps predicted forward) = 972. In addition, the Linear Regression model was trained using log-transformed data and using target values transformed with the Bernoulli link function. Therefore the resulting number of models was 1260. In this section, the main findings from those models are discussed. In order to evaluate model capability of capturing blood glucose patterns, the dataset was divided in two datasets X_{train}, X_{test}, Y_{train} and Y_{test}. For the split, 60% of the data was selected for training and 40% for testing. All the results reported are mean of subject-specific results gathered using the unseen test data.

Overall, the Linear Regression model performed better in the blood glucose forecasting experiments than the fixed size Kernel Regression. Based on the result analysis, the lowest errors were achieved using the 5-minute data for prediction tasks. The models were able to capture relatively well the dynamics of downsampled blood glucose time series. Exemplification visualisation of the ability of the model to predict glucose values can be found in Figures 4.1 and 4.2.

4.2.1 Best results

For the Linear Regression task, the lowest RMSE and highest Pearson and Spearman Correlation for every n-step prediction were achieved using the 5-minute data points.

Overall the best results concerning the number of steps predicted forward were achieved using five-minute data. Highest achieved Pearson correlation was 0.997, the highest Spearman correlation was 0.497, and the lowest RMSE was 0.075. These results were achieved when predicting blood glucose one step forward with 5-minute time intervals. When increasing the number of steps we want to forecast forward, the Pearson and Spearman correlation decreases and RMSE increases proportionally to the number of steps.

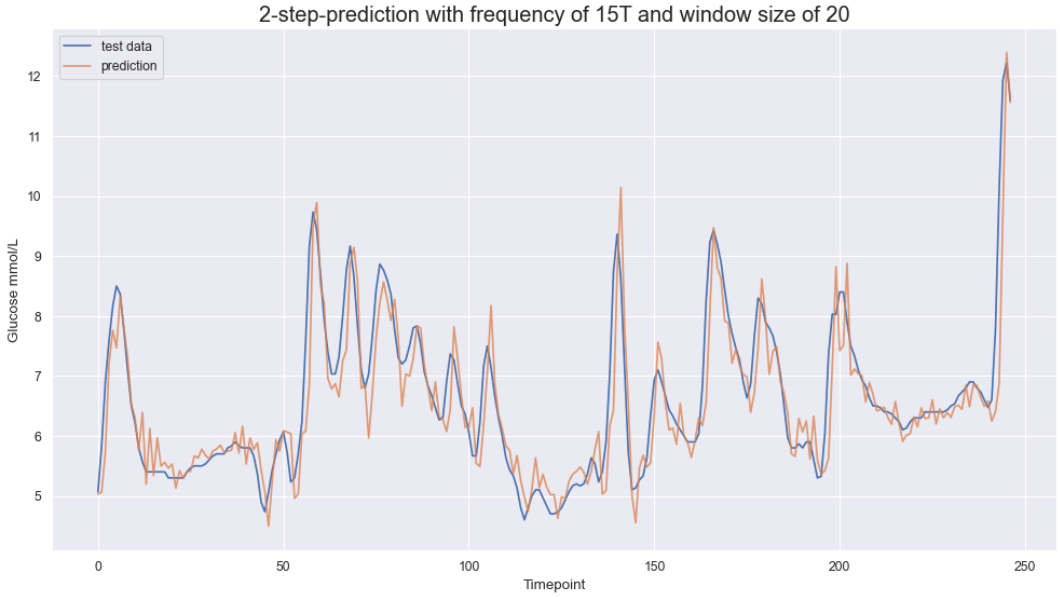

Figure 4.1: Comparison of the test data and model predictions for 2-step forward prediction using 15-minute mean glucose with a window size of 20. Results from Linear Regression model.

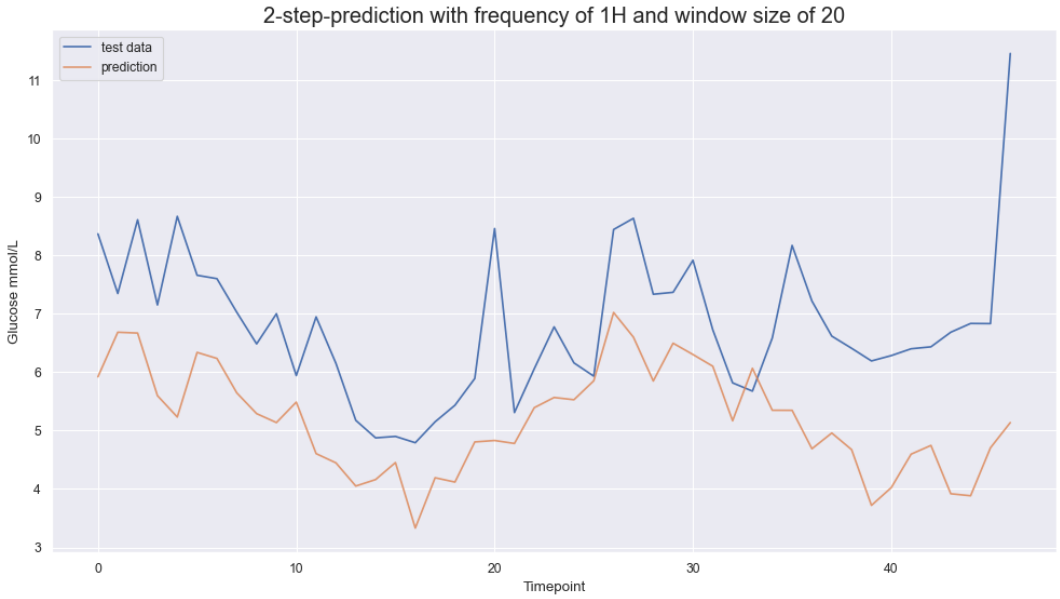

Figure 4.2: Comparison of the test data and model predictions for 2 step forward prediction using 1 hour mean glucose with window size of 20. Results from Linear Regression model.

n-step	w	Pearson
1	3	0.997
2	6	0.990
3	15	0.976
4	10	0.949
5	10	0.911
6	15	0.861

n-step	w	RMSE
1	10	0.075
2	15	0.139
3	15	0.224
4	15	0.323
5	15	0.429
6	15	0.534

n-step	w	Spearman
1	2	0.497
2	6	0.492
3	6	0.483
4	10	0.468
5	6	0.447
6	6	0.422

Table 4.1: Best results for Linear Regression for each step prediction. Run with 5-minute data. n-step is the number of steps predicted forward and w is the window size. (Left: Pearson correlation, Middle: RMSE, Right: Spearman correlation)

n-step	w	Pearson
1	1	0.990
2	1	0.963
3	1	0.921
4	1	0.868
5	1	0.808
6	1	0.745

n-step	w	RMSE
1	1	0.275
2	1	0.394
3	1	0.527
4	1	0.653
5	1	0.772
6	1	0.881

n-step	w	Spearman
1	1	0.494
2	1	0.481
3	1	0.462
4	1	0.438
5	1	0.412
6	1	0.385

Table 4.2: Best results for Kernel Regression with linear kernel for each step prediction. n-step is the number of steps predicted forward and w is the window size. (Left: Pearson correlation, Middle: RMSE, Right: Spearman correlation)

4.2.2 Linear models vs fixed size kernel models

For kernel models, the best-obtained results are presented in the Tables 4.2 and 4.3. Similar to the Linear Regression models, the best results were obtained with 5-minute sample data. For each metric, the best results are better in every n-step prediction task using the basic Linear Regression. The results of the Kernel Regression using polynomial kernel are slightly better than the results of Kernel Regression using a linear kernel.

4.2.3 Effect of window size

To some extent the increasing the window size seems to improve the results. This is expected behaviour due to the seasonality of blood glucose displayed in Chapter 3 Section 3.4. The effect of window size differs when changing the resampling frequency. With five-minute and 15-minute resamplings, increasing the window size lowers the RMSE and increases both correlation

n-step	w	Pearson
1	1	0.973
2	1	0.952
3	1	0.950
4	1	0.873
5	1	0.794
6	1	0.723

n-step	w	RMSE
1	1	0.487
2	1	0.55
3	1	0.628
4	1	0.715
5	1	0.896
6	1	0.969

n-step	w	Spearman
1	1	0.494
2	1	0.481
3	1	0.462
4	1	0.437
5	1	0.416
6	1	0.379

Table 4.3: Best results for Kernel Regression with polynomial kernel for each step prediction. n-step is the number of steps predicted forward and w is the window size. (Left: Pearson correlation, Middle: RMSE, Right: Spearman correlation)

coefficients to some extent. With 30-minute and 1-hour resampling, we can observe an increase in RMSE concerning window size. For 30-sminute data, the correlation coefficients increase with few first increments in window size but decrease when window size increases. For the 1-hour resampling, we observe a similar pattern in RMSE, but the correlation coefficients improve when increasing the window size. Closer inspection of forecasting the glucose value three steps forward with 15-minute time intervals with different window sizes is visible in Figure 4.3. Figure 4.3 shows that model performance increases when increasing the window size from one until after the window size of six; the performance starts to decrease. Similar patterns are observed when using the 5-minute data to forecast glucose values six steps forward. As visible in Figure 4.3 right-hand side, increasing the window size improves the model performance until after the window size of 10, the RMSE stays approximately the same.

4.2.4 Effect of resampling

The effect of resampling is naturally large when comparing the model performance. Changing the resampling frequency changes the task setup from specific glucose value prediction based on earlier values into mean glucose prediction based on earlier glucose means. For example, the 5-step prediction with original 5-minute intervals corresponds to predicting the value after 25 minutes, and the 5-step prediction with 15-minute resampling corresponds to forecasting the 15-minute mean glucose 75 minutes forward. Due to those above, increasing the resampling frequency inevitably decreases the model performance in terms of RMSE, Pearson correlation and Spearman correlation.

The comparison between the effect of resampling in the two-step pre-

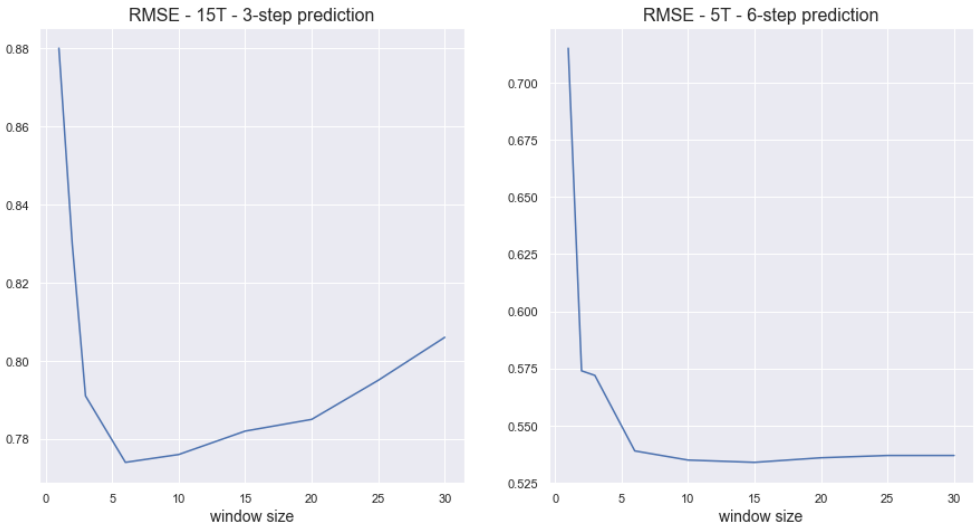

Figure 4.3: Impact of window size in RMSE when using 15-minute resampling with n-step=3 and 5-minute data with n-step=6

diction task is visualised in Figure 4.4. From the figure, we can observe that with 5-minute and 15-minute resampling, increasing the window size seems to increase model performance in terms of RMSE, Pearson correlation and Spearman correlation, to some extent, since there is a slight decrease in performance after a particular window size. For the 30 minute resampling, increasing the window decreases the model RMSE after a window size of 10. Observation of Spearman correlation reveals that the model performance starts decreasing after a window size of three. For the one hour resampling, the performance concerning the RMSE seems to decrease when increasing the window size, until after the window size of 20, the RMSE stabilises and even start to decrease after a window size of 25. The Pearson correlation and Spearman correlation seems to behave differently to the RMSE. The Pearson correlation shows an increase between windows sizes 3-10 and then starts increasing after the window size of 20. Similar patterns but is shown by Spearman correlation, where the performance increased by with window size of 2 and then decreases until after window size of 10 the Spearman correlation starts to increase.

4.2.5 Food input vs without food input

Using the food intake vector had different effect for Linear Regression and Kernel Regression models. In the Linear Regression setup, concatenating the glucose time series data with the food intake binary vector did not affect the

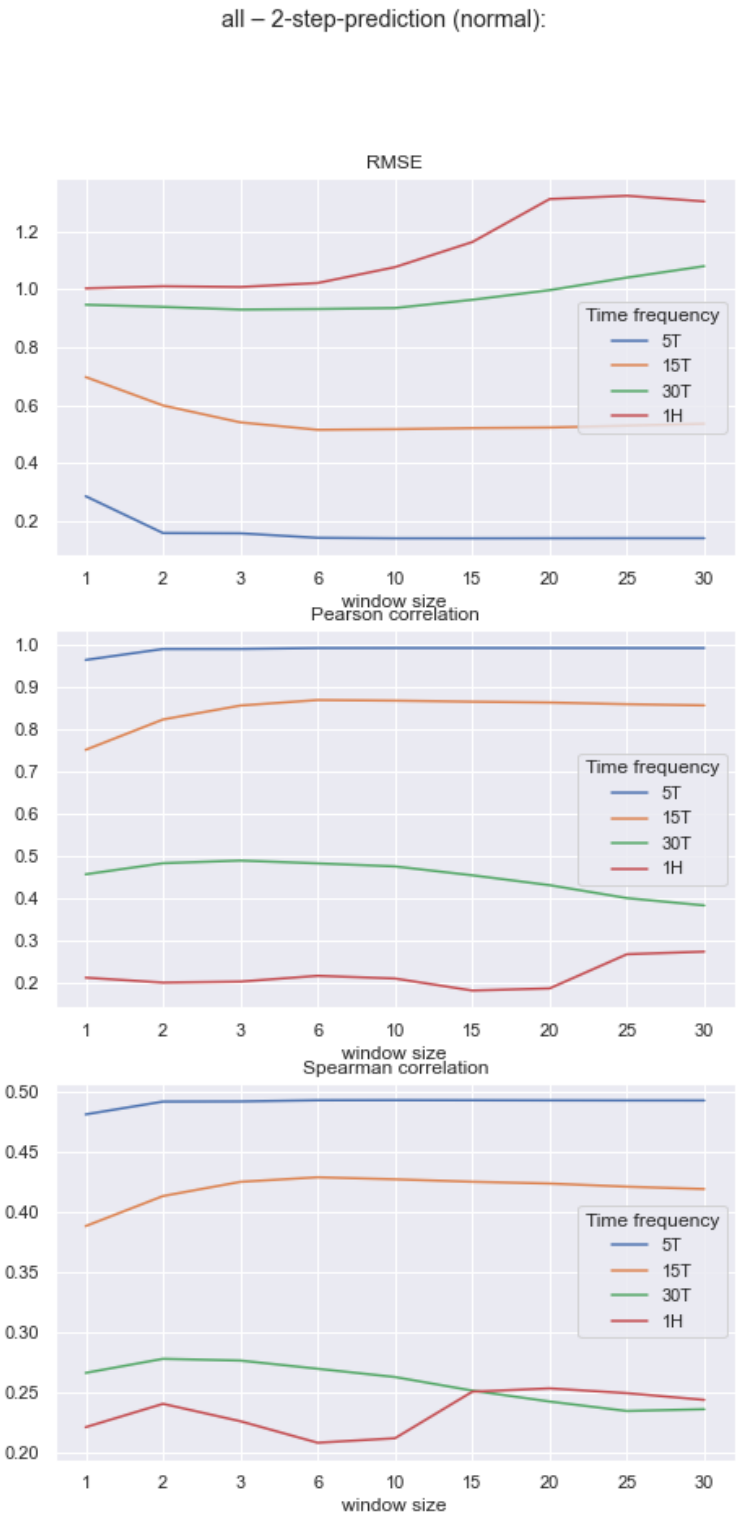

Figure 4.4: Impact of resampling frequency in RMSE, Pearson correlation, Spearman correlation. (n-step=2)

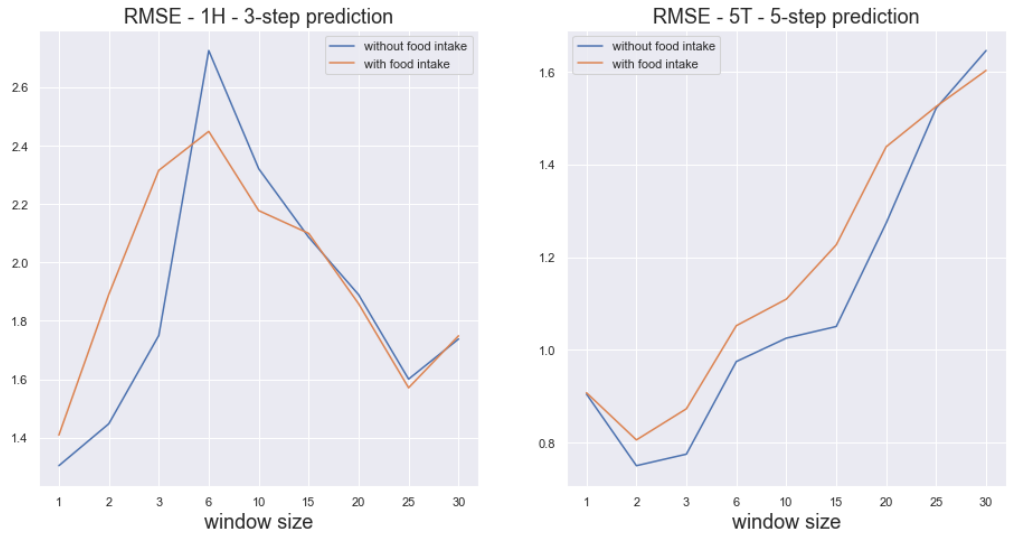

Figure 4.5: Impact of food intake vector in RMSE, when using 5-minute interval data to predict 25 minutes forward (right) and using one hour mean interval data to predict hourly mean after three hours (left).

prediction results notably. In the Kernel Regression setup, taking into account the food intake vector did have a small effect. In short term prediction, for example, predicting glucose values after 25 minutes using five-minute interval data, the model trained with food intake performs worse almost in every time window, excluding window sizes over 25 to 30. The resulting RMSE plots are visible in Figure 4.5. On the other hand, when moving towards longer-term trends, for example, when predicting hourly mean after three hours using one hour mean glucose data, the model with food intake data performs better than the model without the meal input. From the Figure 4.5 we can see that the model with food intake performs better with multiple window sizes, approximately from windows size of 5 to 14 and from 17 to 28. Overall no significant performance improvement was achieved by aggregating the food intake data into the glucose value time series.

4.2.6 Effect of data transformation

Overall, the data transformation did not significantly affect the model performance, providing only a tiny improvement. The effect of data transformation is visualised in Figure 4.6, where Pearson correlations between models using log-transformed, Bernoulli link-function -transformed and normal data are compared when using 15-minute data. As seen from the figure, only minor

improvements are visible, but the learning pattern stays the same, and the best results are observed with the same window size.

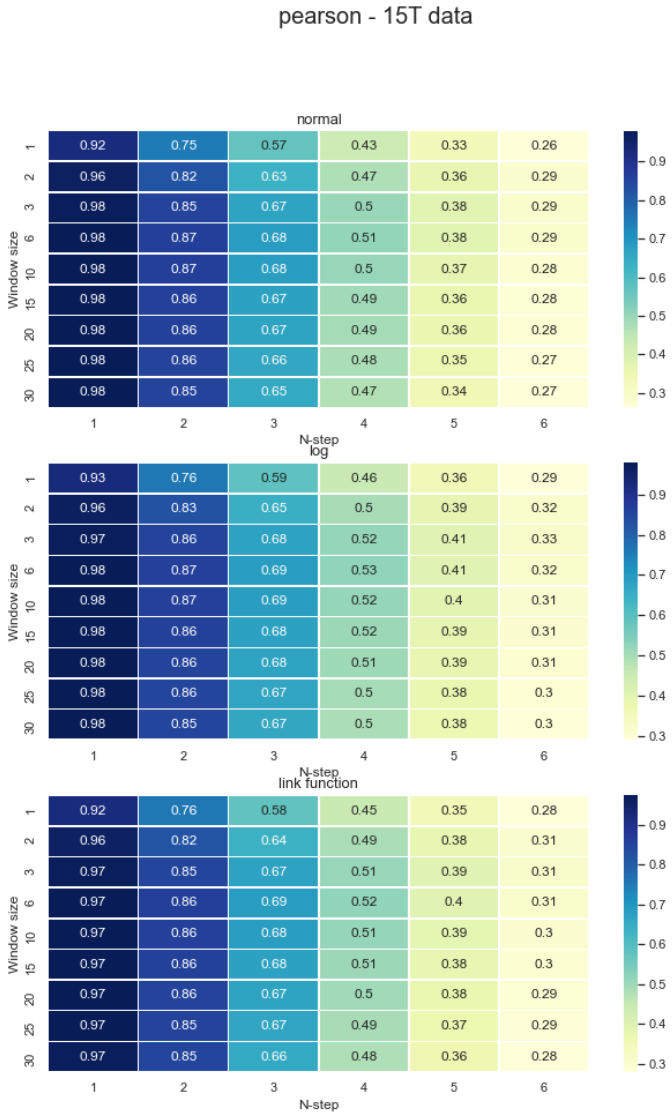

Figure 4.6: Comparison of Pearson correlation between forecasting models with different data preprocessing. Normal refers to not transformed data, original data, log refers to log-transformed data, and link function refers to Bernoulli-link transformed target data. Ran with 15-minute resampling.

Chapter 5

Glucose Profile Applications

After experimenting with the different blood glucose forecasting models, we then proceeded to fit all the models with one week blood glucose time series data and extracted the blood glucose profile vectors as described in Chapter 3. After the extraction process, we had 1260 different blood glucose profiles. The extracted blood glucose profiles were used in two classification applications. The first is T2D/IGT classification, and the second is hyperglycemia prediction. This chapter describes the applications of the extracted blood profile in the applications above.

5.1 Blood glucose profile

Based on the results in Chapter 4, Linear Regression and Kernel Regression were applicable for blood glucose time series forecasting and were able to capture some of the underlying patterns in blood glucose. Following the analogy of study by Plis et al. [42], where the ARIMA model parameters were utilised in blood glucose prediction and hypoglycemia prediction, we hypothesise that the Linear Regression and the Kernel Regression coefficients capture some essential characterisation of subjects blood glucose patterns and therefore can be used in different real-life applications such as T2D/IGT classifications and hyperglycemia prediction.

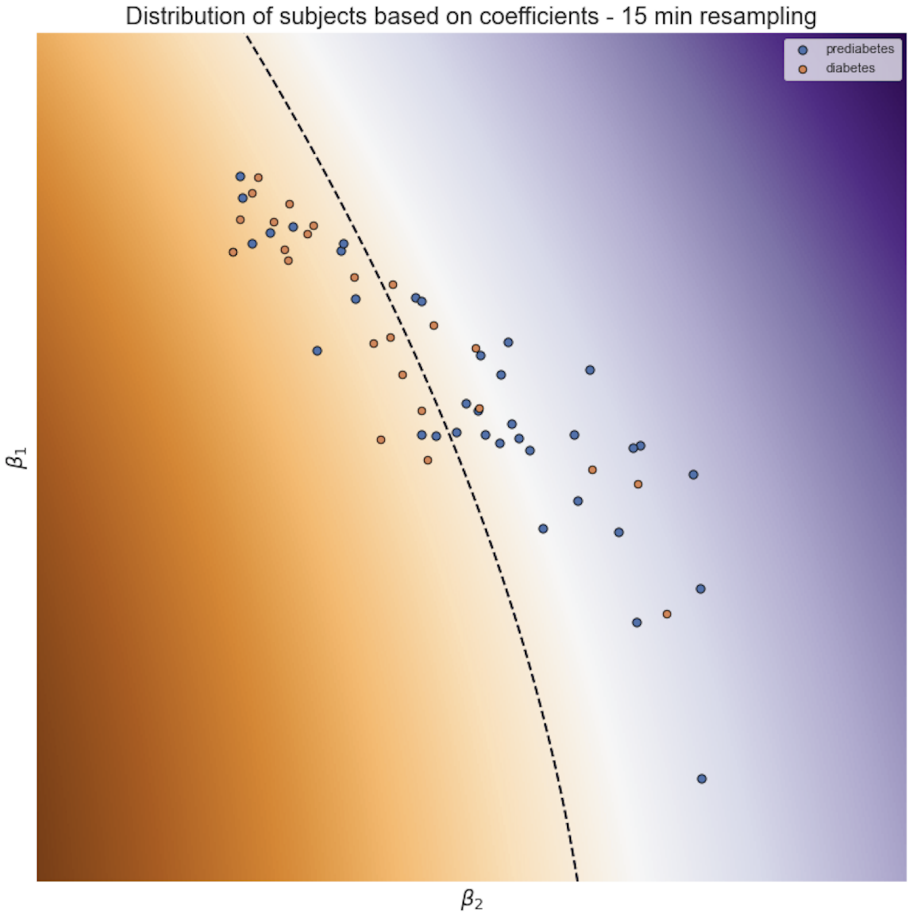

Figure 5.1: Visualisation of the differences in coefficients between diabetic and prediabetic patients after fitting the Linear Regression with window size of two using 15 minute data. The figure is created using SVM to T2D/IGT.

Some support for the characterisation power of coefficients is visible in Figure 5.1. When plotting the coefficient pairs of window size two with 15-minute data for each subject, we can observe that subjects diagnosed with T2D are roughly located on the left top corner, and prediabetics are mostly located near the centre.

5.2 Performance assessment

The comparison between models trained for a classifications task was performed based on *accuracy*, *precision*, and *recall*. In addition, the results were

validated using cross-validation. In this section, we describe the theoretical background of the performance assessment metrics.

5.2.1 Confusion matrix

The confusion matrix is used to visualize the performance of a learning algorithm in a statistical classification problem. The matrix is illustrated in Figure 5.2.

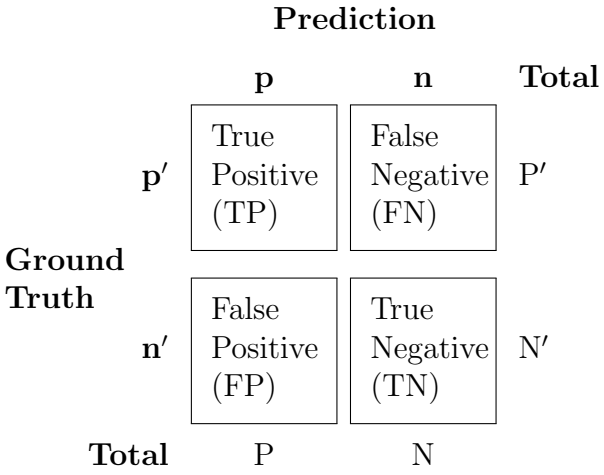

Figure 5.2: Confusion matrix

5.2.2 Accuracy

In the classification domain, accuracy (ACC) measures the amount of correctly classified elements [39]. Accuracy is defined as:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

The main fallacy of using accuracy as a performance metric is that it might not always reveal problems in the classification model. Given the imbalanced dataset with nine target labels that are negative and one positive target label, the model would achieve 90% accuracy by always predicting false labels. Commonly, the performance of classification models is compared using precision and recall.

5.2.3 Precision

Precision, also known as positive predictive value is defined as [39]:

$$Precision = \frac{TP}{TP + FP} \tag{5.2}$$

Intuitively precision can be understood as an amount of relevant predictions among all predictions. Precision describes how many of the instances labelled as positive are truly positive.

5.2.4 Recall

Recall, also known as sensitivity or true positive rate, describes the amount of relevant items classified correctly [39].

Recall is formulated as:

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

For example, in T2D/IGT classification task, recall describes how many T2D subjects the model has detected successfully.

5.2.5 Cross-validation

The performance of supervised machine learning models can be estimated using numerous methods, and one effective method is K-fold cross-validation. The intuition behind the K-fold cross-validation is to avoid the model from overfitting and thus, find more generalised machine learning models [24].

In order to achieve more comparable results in T2D/IGT classification with studies by Acciaroli et al. [1] and Longato et al. [26] and Longato et al. [27], we used the same five-fold cross-validation to assess T2D/IGT classification model performance.

The basic idea of k-fold cross-validation following:

Step 1 Split the dataset D randomly into k number of approximately equal sized and mutually exclusive subsets, also called as folds. This results in a set of folds $S_{k-fold} = \{D_1, D_2 \dots D_k\}$

Step 2 Train the machine learning model k times using always one folds a a test dataset, and the rest folds as a training dataset. Given iteration steps $i \in \{1, 2 \dots k\}$, for each iteration i, we train the machine learning model with dataset $D_{train} = D_i \setminus S_{k-fold}$ and test the performance with dataset D_i

5.3 Diabetes and prediabetes classification

Standard techniques assessing and diagnosing diabetes consists of oral glucose tolerance test (OGTT) and HbA1c values. However, in their study, Madhu et al. [28] showed that analysing glycemic variability indices extracted from CGM-traces could be beneficial in the early identification of subjects with a higher risk of diabetes.

Three different studies, Acciaroli et al. [1], Longato et al. [26] and Longato et al. [27] has used the same dataset in glycemic variability based diabetes and prediabetes classification. The results of these studies are described in Chapter 2 Section 2.3.1.

The first application of the produced blood glucose profile was to use it as a supporting feature in glycemic variability based T2D/IGT classification. IGT is a form of prediabetes, as described in Chapter 2. For the the classification task, we selected to use a subset of glycemic variability indices Acciaroli et al. [1] used in their experiments. The selected features consisted of ten glycemic variability indices: mean glucose, max glucose, min glucose, standard deviation, coefficient of variation, time in range, time above range, time below range, LBGI, HBGI. The used indices are explained in Chapter 3 Section 3.4.

Algorithm description of T2D/IGT classification using glycemic variability indices:

Given n CGM-traces $\{ G_1, G_2 \dots G_n \}$ labeled with class $y \in \{\text{T2D, IGT}\}$:

Step 1 Preprocess the blood glucose time series and event time series based on selection of data preprocessing hyperparameters to the training data $\mathbf{X}$ and the target vector $\mathbf{y}$

Step 2 Fit blood glucose time series forecasting model M with data obtained in Step 1 and extract the blood glucose profile vector $x_{profile}$ for each subject.

Step 3 Extract set of ten glycaemic variability indices from each trace G to form a feature vector x_{GV} containing data of GV indices.

Step 4 Concatenate vectors $x_{profile}$ and x_{GV} to form a feature vector x used in T2D/IGT classification

Step 5 Train the Logistic Regression classifier using stratified five-fold cross validation.

	window size	resample frequency	n-step prediction	Accuracy	Precision	Recall
Benchmark	-	-	-	0.75	0.78	0.625
Best ACC	10	15 min	4	**0.83**	0.82	0.79
Best precision	10	15 min	5	0.83	**0.85**	0.79
Best recall	2	1H	6	0.83	0.79	**0.88**

Table 5.1: Comparison between performance in T2D/IGT classification without and with blood glucose profile. Mean of five-fold cross-validated results using Linear Regression-based blood glucose profile.

5.3.1 T2D/IGT classification results

In the T2D/IGT classification, the blood glucose profile increased the model performance when compared using only the established glycemic variability indices.

Comparison between benchmark model using only glycemic variability indices and models with Linear Regression-based blood glucose profile included is displayed in Table 5.1. The best mean cross-validated accuracy 83% was achieved using the normal data, blood glucose profile fitted with a window size of 10, for four-step prediction using a 15-minute interval glucose time series. The best mean cross-validated precision 85% was achieved with a blood glucose profile fitted with a window size of 10 for five-step prediction using a 15-minute interval glucose time series. The best recall was achieved with a blood glucose profile fitted with a window size of 10 for six-step prediction using a 1-hour interval glucose time series. The best recall was fitted to log-transformed data. The best results were achieved by using models which did not use meal intake as an input.

The best achieved classifications performances with Kernel Regression-based glucose profiles are summarised in Table 5.2. As described in 4 the Kernel Regression task was run with the linear and polynomial kernel, and all the best results were obtained using the linear kernel and using the profile fitted only for the blood glucose data without meal intake info.

	window size	resample frequency	n-step prediction	Accuracy	Precision	Recall
Benchmark	-	-	-	0.75	0.78	0.63
Best ACC	2	15 min	4	**0.85**	0.83	0.79
Best precision	2	15 min	4	0.85	**0.83**	0.79
Best recall	10	5 min	4	0.81	0.75	**0.83**

Table 5.2: Comparison between performance in T2D/IGT classification without and with blood glucose profile. Mean of five-fold cross-validated results using Kernel Regression-based blood glucose profile.

5.4 Hyperglycemia prediction

Hyperglycemia is defined as a state of having a blood glucose level higher than normal. In studies by Acciaroli et al. and Longato et al. [1, 26] the used threshold for hyperglycemia was 180 mg/dL ($\approx$ 10 mmol/l), and the same threshold values were used in this experiment. Predicting future hyperglycemic/hypoglycemic events for 10-30 min ahead are essential, especially for applications for type 1 diabetics, since the effect of insulin occurs within 10–30 min, and the effect of meal response on glucose levels occurs approximately within 5-10 min [49].

Due to the imbalance of hyperglycemic blood glucose values, the class balance between normoglycemic and hyperglycemic values had to be adjusted for the training dataset. Class imbalance visualised in Figure 5.3.. In the training phase, the dataset consists of 25% hyperglycemic values and 75% non-hyperglycemic values. During the test phase, non-balanced glucose data was used to assess model performance in real-life data.

The data for training hyperglycemia predictor was formed as following: Given blood glucose time series $S_x = x_1, x_2 \dots x_n$ measured in time points $\tau_1, \dots, \tau_t, \dots, \tau_m$ with 15 minute interval, window size $w = 6$ and event horizon $h = 3$, the input data matrix X was formed described in Chapter 3 and the target vector $\mathbf{y}_{\mathbf{hyperglycemia}}$ was formed using the following rule:

$$y_i = \begin{cases} 0, & \forall x \in \{x_i, x_i{+}1 \dots x_{i+h}\}, x < 10 \text{ mmol/l} \\ 1, & \text{otherwise} \end{cases} \tag{5.4}$$

High-level description of the hyperglycemia predictor training:

Step 1 Preprocess the blood glucose time series and event time series based on selection of data preprocessing hyperparameters to obtain the training data $\mathbf{X}$ and the target vector $\mathbf{y}$

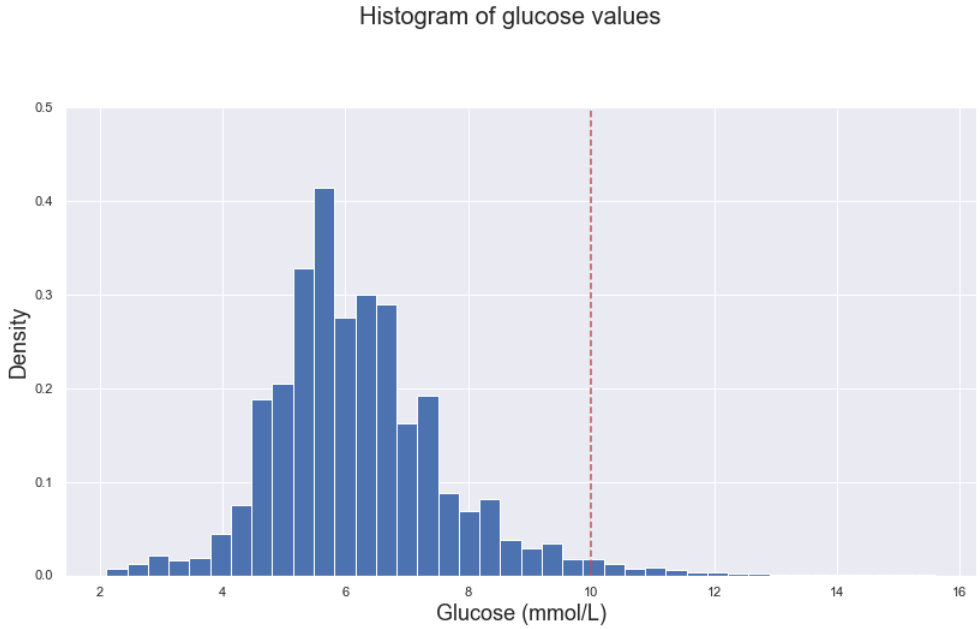

Figure 5.3: Histogram of glucose values in the dataset. Red dashed line corresponds threshold value for hyperglycemia (10 mmol/L)

Step 2 Fit blood glucose time series forecasting model M with data obtained in Step 1 and extract the blood glucose profile vector $x_{profile}$ for each subject.

Step 3 Concatenate each glucose time window vector x_i in $\mathbf{X}$ by corresponding glucose profile vector x_{profile} to form a combined input matrix X_{combined}

Step 3 Train the Logistic Regression model using input matrix X_{combined} and target $\mathbf{y_{hyperglycemia}}$.

5.4.1 Hyperglycemia prediction results

The extracted blood glucose profile had only a minor effect on the model performance in the hyperglycemia prediction task.

The best classification results achieved by utilising the Linear Regression based glucose profile is presented in Table 5.3. The best performance was achieved using a profile extracted from a model run with a window size of 30, a resampling frequency of 30 min, and a one-step prediction. The best results were obtained by extracting the coefficient from a model using non-

	window size	resample frequency	n-step prediction	Accuracy (%)	Precision (%)	Recall (%)
Benchmark	-	-	-	0.94	0.85	0.87
Best accuracy	30	30 min	1	0.94	0.86	0.89
Best precision	30	30 min	1	0.94	0.86	0.89
Best recall	30	30 min	1	0.94	0.86	0.89

Table 5.3: Hyperglycemia prediction using the past glucose values concatenated with blood glucose profile extracted from Linear Regression compared to benchmark

	window size	resample frequency	n-step prediction	Accuracy (%)	Precision (%)	Recall (%)
Benchmark	-	-	-	0.94	0.85	0.87
Best accuracy	25	5 min	5	0.94	0.86	0.88
Best precision	30	30 min	1	0.94	0.87	0.88
Best recall	30	30 min	1	0.94	0.85	0.89

Table 5.4: Hyperglycemia prediction using the past glucose values concatenated with blood glucose profile extracted from Kernel Regression compared to the benchmark.

transformed blood glucose data and using only blood glucose values without a food intake vector.

The best hyperglycemia prediction results achieved for Kernel Regression-based glucose profile are described in Table 5.4. All the results were obtained using the coefficient extracted from Kernel Regression with the linear kernel trained using only blood glucose values. Compared to the best prediction results with the Linear Regression-based profile, there was no significant difference in predictive performance.

Chapter 6

Discussion

As described in Chapter 4, encoding the meal intake event to the blood glucose time series had no significant impact on the predictive performance of the time series forecasting model. Including the meal intake information in the models did still affect the achieved performance, such as decreasing the highest RMSE across all the window sizes. This effect is illustrated in Figure 4.5, where the highest RMSE for the model without food intake is above 2.6. The highest RMSE for the model with food intake is at window size six, and the error is only slightly over 2.4. One reason for the insignificant effect of including the meal intake could be the amount of data. The preprocessing approach for including the food intake described in Chapter 3 where the food intake information is concatenated with blood glucose values might require more data to capture the effect of food. In addition, the effect of meal intake on blood glucose might differ due to numerous factors such as the amount of carbohydrates consumed. Observing Figures A.3 and A.4 in Appendix A reveals that usually there is a visible surge in the blood sugar after a meal, but sometimes blood sugar remains relatively stable. It might also be that the effect of food intake is ingrained in the blood glucose time series. Thus, fitting the blood glucose forecasting model using only blood glucose observations might cover the effect of food intake on an individual's blood glucose behaviour.

Although the best performing models in terms of time series forecasting were the models trained with 5-minute data, the applications revealed that the models trained with different resampling frequencies, that is, 15-minute, 30-minute and one-hour, might capture meaningful characterizations of individual's blood glucose patterns. See Figures 4.1. and 4.2 in Chapter 4. Since the goal of the blood glucose forecasting model was to capture dynamics in individuals blood glucose, the RMSE, Pearson correlation and Spearman correlation did not tell the whole truth about the model ability to achieve that

goal. One way to assess that ability was testing in real-life use cases. Overall, the extracted coefficients from both Linear Regression and Kernel Regression models improved T2D/IGT classification results. The best-achieved accuracy with coefficients included was 85%, the best precision was 85%, and the best recall was 88%. Without coefficients, the model accuracy was 75%, precision was 78%, and recall was 63%, supporting the fact that the extracted coefficients could capture relevant information on blood glucose dynamics. As it can be observed from the Figrue 6.1, the overall accuracy when using coefficients extracted from a model fitted with 5-minute data was worse than when using model coefficients trained with 15-minute, 30-minute and one-hour data. This observation could indicate that significant differences in blood glucose patterns between IGT and T2D subjects are mostly visible in longer-term patterns. In addition, the discrepancy might be since the forecasting tasks with 5-minute data might be easier to learn since the forecasting time horizon is shorter than the forecasting tasks with other experimented resampling frequencies. Another reason could be that the model might have the same performance in forecasting tasks with 5-minute data despite the subject's physiological state. Therefore the model is not capturing any difference between the inter-individual blood glucose dynamics. Further research is required to validate the correlation between the regression coefficients and individuals blood glucose dynamics.

In the hyperglycemic event prediction, the blood glucose profile did not improve the model performance significantly. There was no significant difference between the coefficient extracted from the Kernel Regression and the Linear Regression. The accuracy stayed in the 94% in all models, and the best-achieved precision was 87%, which is slightly better than the benchmark result without coefficients. In terms of recall, the best result with coefficients was 89%, indicating only a slight increase compared to the benchmark model result 87%. Based on the results, more research is needed for assessing the usability of the regression coefficients in the hyperglycemia prediction task.

6.1 Recommendation of future studies

The implementation in the thesis did not consider the food nutrients related to the meal but treated the events only as a binary, that is, either the particular event is started at the particular time point or not. Using more information on the food nutrients such as carbohydrate intake used in the study by Plis et al. [42] could improve the blood glucose profile characterisation. The model could also be improved by attaching the exercise event information on the blood glucose model, including exercise start time,

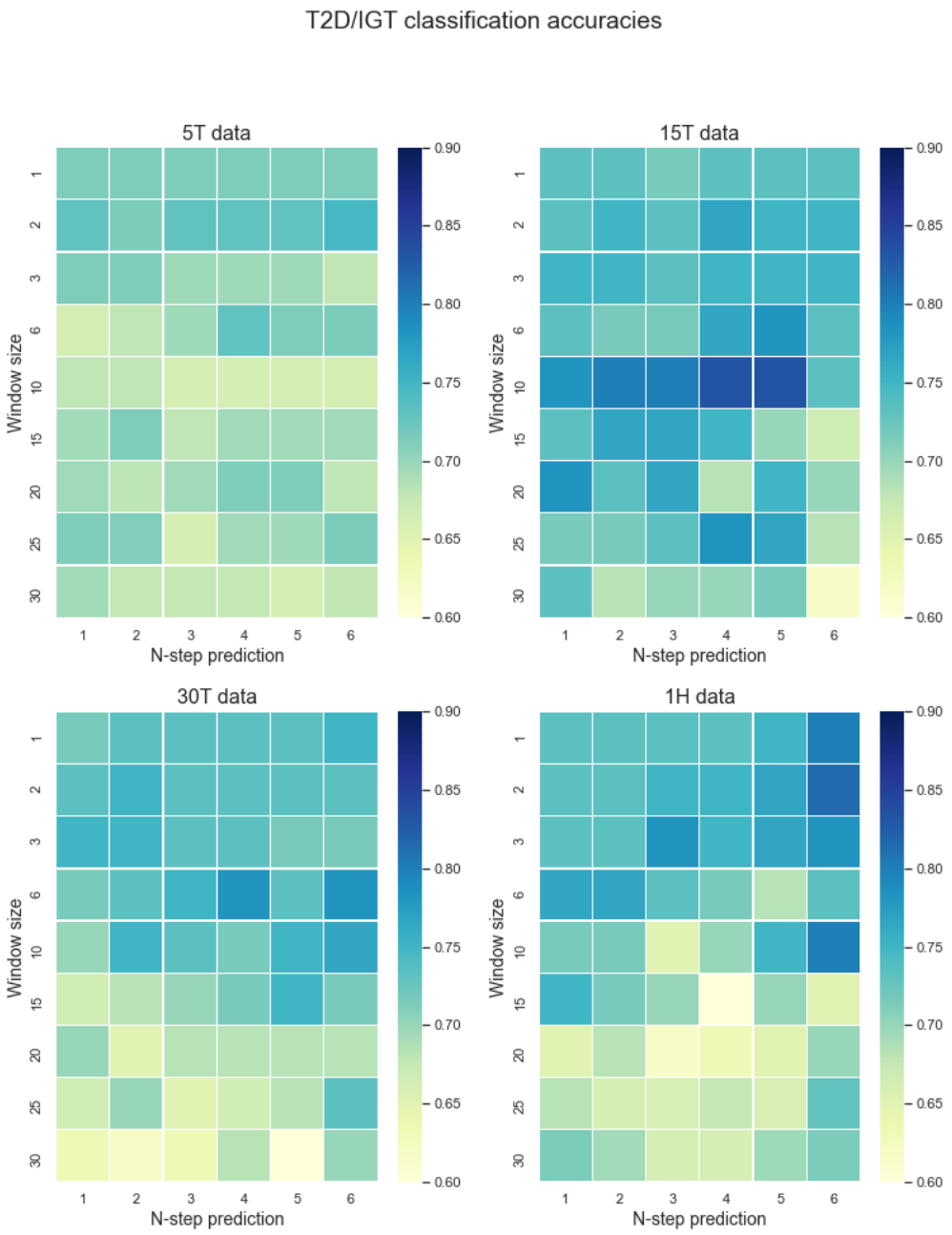

Figure 6.1: Accuracies achieved in T2D/IGT classification using coefficients from different blood glucose forecasting models. Each grid represents accuracies achieved using coefficients extracted from different resampling frequency. The vertical axis represents the selected window size, and the horizontal axis represents the selected number of steps predicted forward.

duration, and heart rate during an exercise.

In future studies, the difference in blood glucose behaviour between weekdays would be interesting to consider, especially when trying to capture latent structure in the blood glucose time series. As displayed in Figure 3.2 in Chapter 3, the average standard deviation was highest on Saturday, indicating that taking into account the weekday might be relevant when developing a blood glucose forecasting model.

Further research could focus on the more in-depth analysis of the impact of blood glucose profile in blood glucose dysregulation assessment. The extracted coefficients could be compared with other glycemic variability indices to find possible correlations between the indices and the coefficients. In addition, the applicability of the extracted blood glucose vector could be tested with other machine learning classifications techniques such as SVM and neural network.

The discrepancy between the best predictive models and the usefulness of the extracted regression coefficients in medical applications is also something that future research might focus on. It would be interesting to know why the extracted coefficients from the model with the best forecasting accuracy were not the best to separate subjects between T2D and prediabetes and why the regression coefficients trained with 15-minute frequency time series improved the classification performance the most.

The resulting cross-validated classification accuracies when using the regression coefficients and glycemic variability indices were better than cross-validated accuracies Acciaroli et al. [1], Longato et al. [26] and Longato et al. [27] reported in their studies. Despite the better results, more research is still needed. The approach described in this thesis and the approaches used in the earlier studies mentioned above should be tested in the same experimental setting and assessed using the same performance assessment methods and metrics to have more comparable results.

The blood glucose characterization could also be investigated more in online-learning applications. One could investigate how the blood glucose characterisation vector changes over time and is it possible to capture an individual's physiological state via these characterisations, or could the characterization be utilized in artificial pancreas applications described in Chapter 2.

6.2 Limitations of the study

In this thesis, we used relatively simple methods for extracting the blood glucose characterisation vector and to assess the suitability of the blood glucose

vector in real-life use cases. No broad hyperparameters search was performed for the glucose forecasting models or classification models, leaving room for optimisation in forecasting and classification performance. As described in Chapter 2 there are numerous factors affecting individuals blood sugar in addition to those used in this thesis experiments, that is, blood glucose and meal intake.

When assessing the model performance in one population, one should be careful when assuming generalisation to other populations, such as the population of different age. For example, Barakat et al. [4] found that cutoff values extracted from an SVM for a diabetes diagnosis trained with data from the Oman population differed from American Diabetes Association and WHO guideline values for diagnosing diabetes. As described in Chapter 3 the dataset contains only data from subjects aged from 44 to 75 with either IGT or T2D, thus representing only a particular population. In addition, the dataset was relatively small, containing only data of 62 subjects. Due to the limited amount of data, further research is required to compare the achieved results with a more representative dataset.

One limitation of the study is the accuracy of CGM. As described in Chapter 2 the readings from CGM are lagging the actual blood glucose concentrations approximately 10-15 minutes. This lag causes a problem since, in some blood glucose prediction settings, it might be essential to know results near real-time to take corrective actions, such as injecting the right amount of insulin. Another problem is raised since the values provided by the CGM device might differ from those measured directly from the blood. For example, as discussed in Chapter 2, the values might be lower when measuring glucose with a CGM than POC. The CGM data used in this thesis was collected with Guardian Real Time or the iPro CGM systems, which are calibrated with actual blood glucose values to reduce the error. An error might still occur due to the human factor during the calibration process. One could use different methods to reduce error, such as Kalman Filtering used, for example, in the patent of Diagnostics Operations Inc. [11], but such methods would have required a larger dataset.

6.3 Conclusions

This thesis explored different blood glucose dynamics modelling approaches and explored if time series forecasting models can capture clinically meaningful blood glucose behaviour.

In the experiment setup, Linear Regression performed better in predicting blood glucose values than the Kernel Regression. Still, both methods

were able to capture patterns in the blood glucose data. Best results in RMSE, Pearson correlation and Spearman correlation were achieved using five-minute data for short term blood glucose prediction. Including food intake data did not have notable effect on blood glucose forecasting accuracy. Despite the best performance in the blood glucose prediction, the coefficients from models trained with five-minute data were not the best features when utilised in classifying T2D/IGT subjects or hyperglycemia prediction.

Based on the T2D/IGT classification experiments, the extracted coefficients improved the model performance, indicating that the models might capture functional characterisation of individuals' blood glucose dynamics. Best cross-validated accuracy 85 % was achieved using coefficients from the Kernel Regression task; the best cross-validated precision 85 % and recall 88% was achieved with Linear Regression coefficients. In the hyperglycemia prediction task, the blood glucose profile did not have a notable impact.

The thesis suggests that regression coefficients could capture meaningful information about individuals' blood glucose dynamics. Given the limitations of the research setting, the achieved results should be considered preliminary. Further research and clinical validation are required due to the small sample size and possible inaccuracies in the CGM technology. Furthermore, the achieved results should be compared extensively with current medically validated methods for assessing blood glucose dynamics.

Bibliography

[1] Acciaroli, G., Sparacino, G., Hakaste, L., Facchinetti, A., Di Nunzio, G. M., Palombit, A., Tuomi, T., Gabriel, R., Aranda, J., Vega, S., et al. Diabetes and prediabetes classification using glycemic variability indices from continuous glucose monitoring data. *Journal of diabetes science and technology 12*, 1 (2018), 105–113.

[2] Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The lancet Diabetes & endocrinology 6*, 5 (2018), 361–369.

[3] American Diabetes Association. Good to know: Factors affecting blood glucose. *Clinical Diabetes 36*, 2 (2018), 202–202.

[4] Barakat, N., Bradley, A. P., and Barakat, M. N. H. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine 14*, 4 (2010), 1114–1120.

[5] Bazaev, N., Pletenev, A., and Pozhar, K. Classification of factors affecting blood glucose concentration dynamics. *Biomedical Engineering 47*, 2 (2013), 100–103.

[6] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238* (2013).

[7] Bunescu, R., Struble, N., Marling, C., Shubrook, J., and Schwartz, F. Blood glucose level prediction using physiological models

and support vector regression. In *2013 12th International Conference on Machine Learning and Applications* (2013), vol. 1, IEEE, pp. 135–140.

[8] Centers for Disease Control and Prevention. Prediabetes - your chance to prevent type 2 diabetes, 2020. `https://www.cdc.gov/diabetes/basics/prediabetes.html` Accessed 23.3.2021.

[9] Daskalaki, E., Diem, P., and Mougiakakou, S. G. Model-free machine learning in biomedicine: Feasibility study in type 1 diabetes. *PloS one 11*, 7 (2016), e0158722.

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

[11] Duke, D. L., Soni, A. S., and Weinert, S. Methods and systems for processing glucose data measured from a person having diabetes. US8843321, 2014. `https://www.freepatentsonline.com/8843321.html` Accessed 18.2.2021.

[12] Eljil, K. S., Qadah, G., and Pasquier, M. Predicting hypoglycemia in diabetic patients using time-sensitive artificial neural networks. *International Journal of Healthcare Information Systems and Informatics (IJHISI) 11*, 4 (2016), 70–88.

[13] Funtanilla, V. D., Caliendo, T., and Hilas, O. Continuous glucose monitoring: a review of available systems. *Pharmacy and Therapeutics 44*, 9 (2019), 550.

[14] Galindo, R. J., Migdal, A. L., Davis, G. M., Urrutia, M. A., Albury, B., Zambrano, C., Vellanki, P., Pasquel, F. J., Fayfman, M., Peng, L., et al. Comparison of the freestyle libre pro flash continuous glucose monitoring (cgm) system and point-of-care capillary glucose testing in hospitalized patients with type 2 diabetes treated with basal-bolus insulin regimen. *Diabetes Care 43*, 11 (2020), 2730–2735.

[15] Georga, E. I., Protopappas, V. C., Ardigo, D., Polyzos, D., and Fotiadis, D. I. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. *Diabetes technology & therapeutics 15*, 8 (2013), 634–643.

[16] Hall, H., Perelman, D., Breschi, A., Limcaoco, P., Kellogg, R., McLaughlin, T., and Snyder, M. Glucotypes reveal new patterns of glucose dysregulation. *PLoS biology 16*, 7 (2018), e2005143.

[17] Heller, A., and Feldman, B. Electrochemistry in diabetes management. *Accounts of chemical research 43*, 7 (2010), 963–973.

[18] Hirsch, I. B. Glycemic variability and diabetes complications: does it matter? of course it does! *Diabetes care 38*, 8 (2015), 1610–1614.

[19] Hogg, R. V., McKean, J., and Craig, A. T. *Introduction to mathematical statistics*. Pearson Education, 2005.

[20] Hyndman, R., and Athanasopoulos, G. *Forecasting: Principles and Practice*, 2nd ed. OTexts, Australia, 2018.

[21] Inayama, Y., Yamanoi, K., Shitanaka, S., Ogura, J., Ohara, T., Sakai, M., Suzuki, H., Kishimoto, I., Tsunenari, T., and Suginami, K. A novel classification of glucose profile in pregnancy based on continuous glucose monitoring data. *Journal of Obstetrics and Gynaecology Research* (2021).

[22] Jensen, M. H., Christensen, T. F., Tarnow, L., Seto, E., Dencker Johansen, M., and Hejlesen, O. K. Real-time hypoglycemia detection from continuous glucose monitoring data of subjects with type 1 diabetes. *Diabetes technology & therapeutics 15*, 7 (2013), 538–543.

[23] Jung, H. S. Clinical implications of glucose variability: chronic complications of diabetes. *Endocrinology and Metabolism 30*, 2 (2015), 167.

[24] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[25] Kovatchev, B. P., Shields, D., and Breton, M. Graphical and numerical evaluation of continuous glucose sensing time lag. *Diabetes technology & therapeutics 11*, 3 (2009), 139–143.

[26] Longato, E., Acciaroli, G., Facchinetti, A., Hakaste, L., Tuomi, T., Maran, A., and Sparacino, G. Glycaemic variability-based classification of impaired glucose tolerance vs. type 2 diabetes using continuous glucose monitoring data. *Computers in biology and medicine 96* (2018), 141–146.

[27] Longato, E., Acciaroli, G., Facchinetti, A., Maran, A., and Sparacino, G. Simple linear support vector machine classifier can distinguish impaired glucose tolerance versus type 2 diabetes using a reduced set of cgm-based glycemic variability indices. *Journal of diabetes science and technology 14*, 2 (2020), 297–302.

[28] Madhu, S. V., Muduli, S. K., and Avasthi, R. Abnormal glycemic profiles by cgms in obese first-degree relatives of type 2 diabetes mellitus patients. *Diabetes technology & therapeutics 15*, 6 (2013), 461–465.

[29] Marathe, P. H., Gao, H. X., and Close, K. L. American d iabetes a ssociation s tandards of m edical c are in d iabetes 2017, 2017.

[30] Marcus, Y., Eldor, R., Yaron, M., Shaklai, S., Ish-Shalom, M., Shefer, G., Stern, N., Golan, N., Dvir, A. Z., Pele, O., et al. Improving blood glucose level predictability using machine learning. *Diabetes/metabolism research and reviews 36*, 8.

[31] Marling, C., and Bunescu, R. C. The ohiot1dm dataset for blood glucose level prediction. In *KHD@ IJCAI* (2018).

[32] Matabuena, M., Félix, P., Meijide-Garcia, C., and Gude, F. Glucose values prediction five years ahead with a new framework of missing responses in reproducing kernel hilbert spaces, and the use of continuous glucose monitoring technology. *arXiv preprint arXiv:2012.06564* (2020).

[33] Medtronic. Sensors transmitters - calibrating your sensor, 2021. `https://www.medtronicdiabetes.com/customer-support/sensors-and-transmitters-support/calibration-sensor` Accessed 17.5.2021.

[34] Mhaskar, H. N., Pereverzyev, S. V., and van der Walt, M. D. A deep learning approach to diabetic blood glucose prediction. *Frontiers in Applied Mathematics and Statistics 3* (2017), 14.

[35] Mustajoki, Pertti, b. . L. p. . K. y. p. n. *Painoindeksi (BMI).*

[36] Nalysnyk, L., Hernandez-Medina, M., and Krishnarajah, G. Glycaemic variability and complications in patients with diabetes mellitus: evidence from a systematic review of the literature. *Diabetes, Obesity and Metabolism 12*, 4 (2010), 288–298.

[37] Nathan, D. M., Davidson, M. B., DeFronzo, R. A., Heine, R. J., Henry, R. R., Pratley, R., and Zinman, B. Impaired fasting glucose and impaired glucose tolerance: implications for care. *Diabetes care 30*, 3 (2007), 753–759.

[38] Noaro, G., Cappon, G., Vettoretti, M., Sparacino, G., Del Favero, S., and Facchinetti, A. Machine-learning based model to improve insulin bolus calculation in type 1 diabetes therapy. *IEEE Transactions on Biomedical Engineering 68*, 1 (2020), 247–255.

[39] Olson, D. L., and Delen, D. *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[40] Pappada, S. M., Cameron, B. D., Rosman, P. M., Bourey, R. E., Papadimos, T. J., Olorunto, W., and Borst, M. J. Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes technology & therapeutics 13*, 2 (2011), 135–141.

[41] Perveen, S., Shahbaz, M., Keshavjee, K., and Guergachi, A. Prognostic modeling and prevention of diabetes using machine learning technique. *Scientific reports 9*, 1 (2019), 1–9.

[42] Plis, K., Bunescu, R., Marling, C., Shubrook, J., and Schwartz, F. A machine learning approach to predicting blood glucose levels for diabetes management. In *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence* (2014).

[43] Rao, K. R., and Yip, P. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[44] Rao, S. S., Disraeli, P., and McGregor, T. Impaired glucose tolerance and impaired fasting glucose. *American family physician 69*, 8 (2004), 1961–1968.

[45] Saiti, K., Macaš, M., Štechová, K., Pit'hová, P., and Lhotská, L. Predicting blood glucose levels for a type i diabetes patient by combination of autoregressive with one compartment open model. In *EMBEC & NBC 2017*. Springer, 2017, pp. 771–774.

[46] Scanlon, V. C., and Sanders, T. Essentials of anatomy and physiology fifth edition.

[47] Seo, W., Lee, Y.-B., Lee, S., Jin, S.-M., and Park, S.-M. A machine-learning approach to predict postprandial hypoglycemia. *BMC medical informatics and decision making 19*, 1 (2019), 1–13.

[48] Shalev-Shwartz, S., and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[49] Snetselaar, L. G. *Nutrition counseling skills for the nutrition care process*. Jones & Bartlett Learning, 2009.

[50] Sprent, P. *Applied nonparametric statistical methods*. Springer Science & Business Media, 2012.

[51] Sutton, R., and Barto, A. *Reinforcement Learning: An Introduction*, second ed. MIT Press, 2018.

[52] Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., and Kivimäki, M. Prediabetes: a high-risk state for developing diabetes. *Lancet 379*, 9833 (2012), 2279.

[53] Tejedor, M., Woldaregay, A. Z., and Godtliebsen, F. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artificial Intelligence in Medicine* (2020), 101836.

[54] Wang, J. Electrochemical glucose biosensors. *Chemical reviews 108*, 2 (2008), 814–825.

[55] Welling, M. Kernel ridge regression. *Max Welling's Classnotes in Machine Learning* (2013), 1–3.

[56] Wiley, M. T. *Machine learning for diabetes decision support*. PhD thesis, Ohio University, 2011.

[57] Woldaregay, A. Z., Årsand, E., Botsis, T., Albers, D., Mamykina, L., and Hartvigsen, G. Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes. *Journal of medical Internet research 21*, 5 (2019), e11030.

[58] Xie, J., and Wang, Q. Benchmark machine learning approaches with classical time series approaches on the blood glucose level prediction challenge. In *KHD@ IJCAI* (2018).

Appendix A

Blood Glucose Graphs

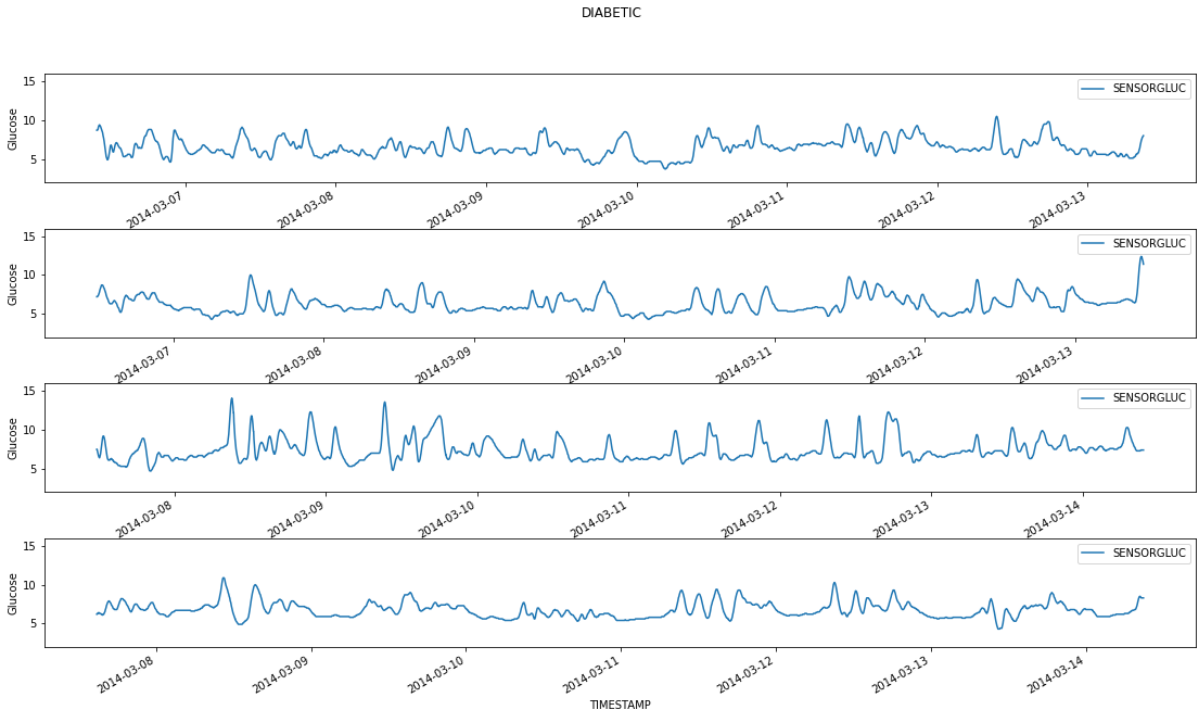

Figure A.1: Seven days glucose time series plots of four diabetic subjects.

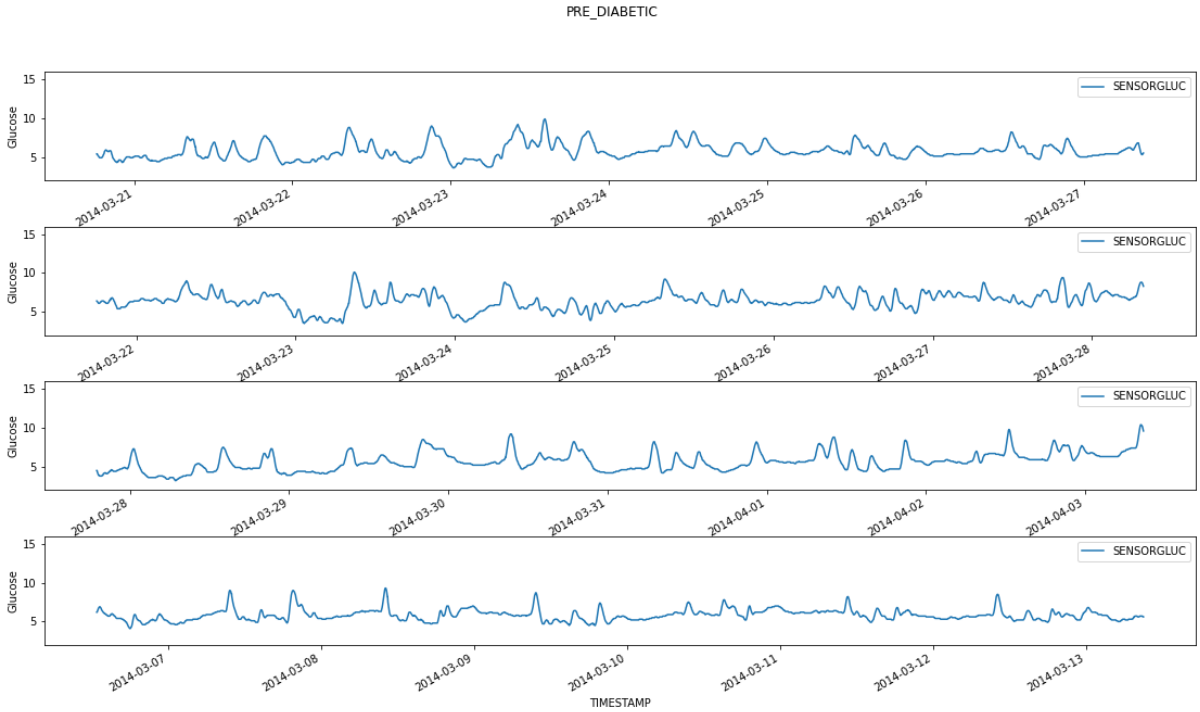

Figure A.2: Seven days glucose time series plots of four prediabetic subjects.

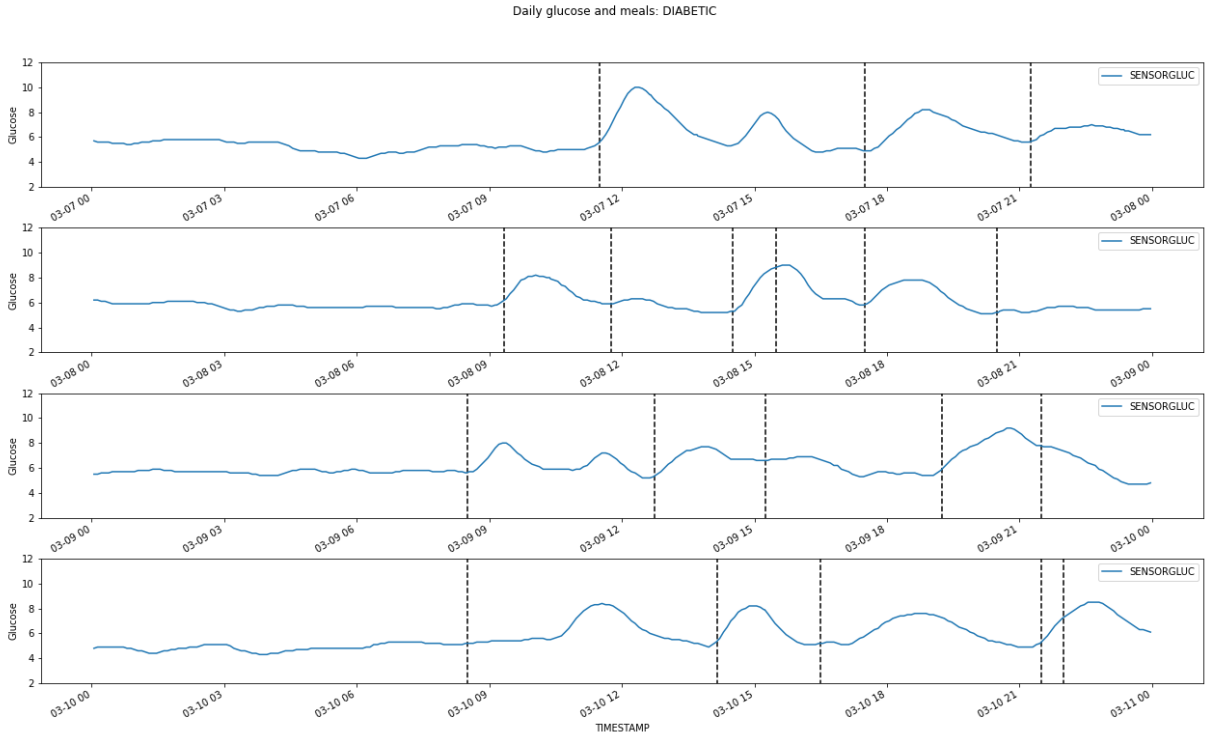

Figure A.3: Daily glucose time series plots of randomly selected diabetic patients with meals plotted as dashed vertical line.

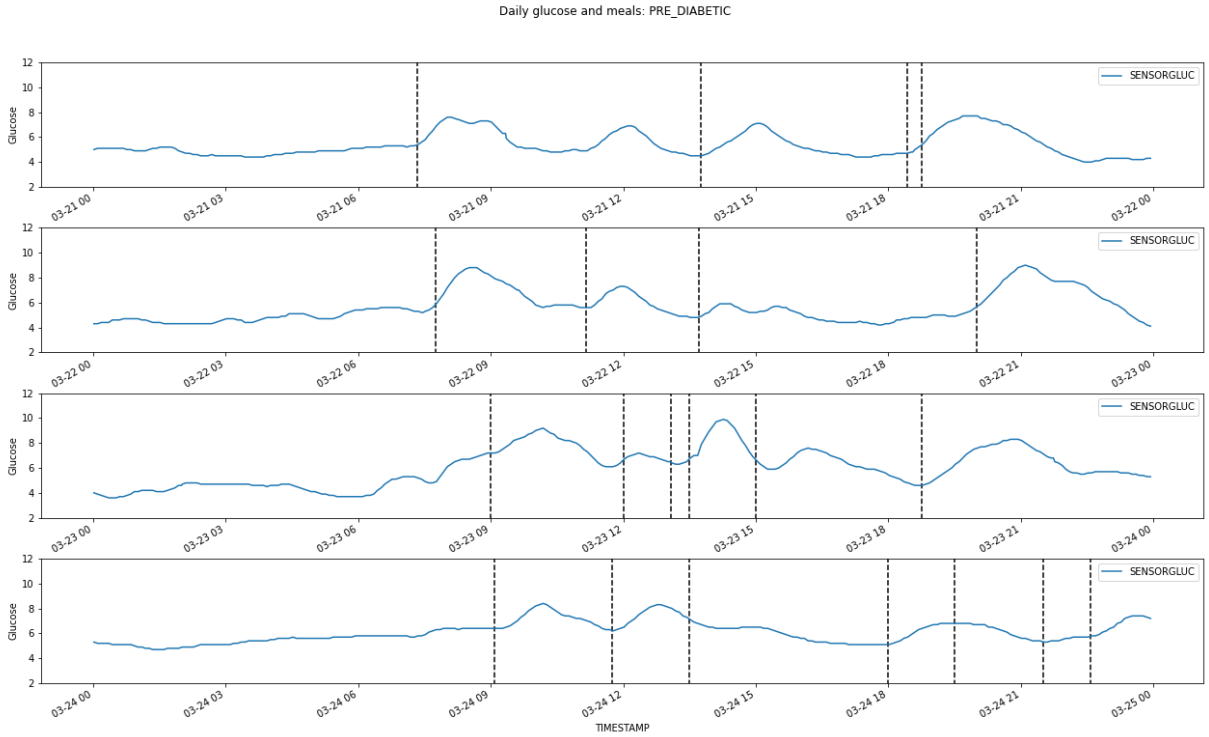

Figure A.4: Daily glucose time series plots of randomly selected prediabetic patients with meals plotted as dashed vertical line