

Imperial College of Science, Technology and Medicine
Department of Computing

Machine Learning Algorithms for Blood Glucose Prediction in Diabetes Management

Ioannis Konstantinos Afentakis

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Artificial Intelligence and Machine Learning
Imperial College London, March 20, 2025

Statement of Originality

I hereby declare that the content presented in this thesis is the result of my own work under the supervision of Prof. Pantelis Georgiou and Prof. Nick Oliver. All sources of information and ideas derived from the work of others have been acknowledged and references have been provided where relevant. Additionally, any data used, either for models development or for validation, have also been acknowledged accordingly.

Ioannis Konstantinos Afentakis

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Ioannis Konstantinos Afentakis

Abstract

People with type 1 diabetes (T1D) depend on externally administered insulin to regulate their blood glucose levels and prevent long-term complications. Managing diabetes involves a complex set of tasks, including glucose monitoring, insulin administration, carbohydrate counting, and maintaining a regular exercise routine, among other lifestyle choices. Achieving the recommended glycemic target is challenging, with only about one out of four individuals successfully managing their condition.

Clinical wearable devices, such as continuous glucose monitors, insulin pumps and activity trackers, play a crucial role in helping people with T1D to control their blood glucose levels. Artificial intelligence (AI) and decision support systems that leverage the wealth of data generated by these devices can offer significant assistance in managing T1D effectively and enhancing the overall quality of life of those living with the condition.

This thesis focuses on building Machine Learning (ML) models using glucose, insulin, meal, and activity data, to predict potential future adverse events, such as nocturnal hypoglycaemia, recommend carbohydrate intake to prevent low glucose levels, or forecast forthcoming blood glucose values. These models can be used as the backbone of an AI decision support system for T1D management outside of clinical settings, mitigating the risk of health complications and the progressive deterioration of the disease.

Acknowledgements

First and foremost I would like to express my gratitude to my supervisor Prof. Pantelis Georgiou for giving me the opportunity to work on this research project. His guidance, support and encouragement during the last few years was endless and invaluable.

I am very grateful to my clinical supervisor Prof. Nick Oliver. His expertise and insights allowed me to better understand the challenges of diabetes management and explore new ideas and opportunities in this exciting research field.

I am thankful to Dr. Pau Herrero for sharing his knowledge and guidance in the beginning of my research. His ideas and suggestions play a crucial role in the direction of this research project. Special thanks also go to my colleagues for the great collaboration and fruitful discussions.

I would also like to acknowledge the AI4Health CDT program at Imperial College London for funding my research, as well as all the patients and people with diabetes for giving consent and sharing their data for research purposes. Without them this and many other studies would not have been possible.

I acknowledge the use of ChatGPT 3.5 (OpenAI, <https://chat.openai.com/>) to assist in rephrasing and refining content that I originally created and provided as input. I confirm that no content generated by AI has been presented as my own work.

Last but not least, I would like to express my gratitude to my parents who always believe in me and support me.

Dedication

Dedicated to my parents.

Contents

Statement of Originality	i
Copyright Declaration	iii
Abstract	v
Acknowledgements	vii
1 Introduction	1
1.1 Motivation	1
1.2 Research Objectives	3
1.3 Thesis Structure	4
1.4 Thesis Contributions	5
2 Background	7
2.1 Diabetes Mellitus	7
2.2 Challenges of Diabetes Management	9
2.3 Fundamentals of Machine Learning	11

2.3.1	Supervised Machine Learning	11
2.3.2	Unsupervised Machine Learning	12
2.4	Review of Blood Glucose Level Prediction	13
2.5	Review of Nocturnal Hypoglycaemia Prediction	15
2.6	Review of Decision Support Systems in Diabetes Management	18
2.7	Conclusion	19
3	Blood Glucose Level Prediction	20
3.1	Introduction	20
3.2	Data	21
3.2.1	OhioT1DM dataset	21
3.2.2	GVAS dataset	22
3.3	Algorithms and Model Development	23
3.3.1	ARIMA(p,d,q)	24
3.3.2	RNN LSTM	25
3.4	Results and Discussion	27
3.5	Conclusion	37
4	Prediction of Nocturnal Hypoglycaemia	38
4.1	Introduction	38
4.2	Data	39
4.2.1	ABC4D dataset	39

4.2.2	IHART dataset	39
4.3	Algorithms	40
4.3.1	Random Forests	41
4.3.2	Support Vector Machines (SVM)	42
4.4	Model Development	45
4.4.1	Data Preprocessing	45
4.4.2	Feature Extraction	45
4.4.3	Training and Validation	46
4.5	Results	49
4.6	Discussion	51
4.7	Conclusion	55
5	Applications and Clinical Trial Design	56
5.1	Introduction	56
5.2	In-silico study	56
5.2.1	T1DM Simulator	56
5.2.2	Study Design	59
5.2.3	Results and Discussion	61
5.3	Fundamentals of Clinical Trials	64
5.3.1	Phases of Clinical Trials	65
5.3.2	Types of Clinical Trial Designs	65
5.3.3	Randomization and Blinding	68

5.4	Towards a Clinical Trial for NH Prevention	69
5.4.1	Decision Support System Design	69
5.4.2	Clinical Trial Design and Discussion	71
5.5	Conclusion	73
6	Conclusion	75
6.1	Summary of Results	75
6.2	Thesis Achievements	76
6.3	Thesis Limitations and Future Work	77
A	Supplementary Materials	82
	Bibliography	82

List of Tables

2.1	Summary of Previous Studies from the KDD BGLP Challenge using variations of LSTM models along with performance in 30 and 60 minutes prediction horizon.	14
2.2	Summary of Previous Studies on Nocturnal Hypoglycaemia Prediction.	16
3.1	Demographic and clinical characteristics ($Mean \pm SD$) of the subjects in the two clinical datasets.	23
3.2	Personalised models performance with 30 minutes prediction horizon.	28
3.3	Personalised models performance with 60 minutes prediction horizon.	28
3.4	Personalised models performance with 90 minutes prediction horizon.	29
3.5	Personalised models performance with 120 minutes prediction horizon.	29
3.6	Prediction Time Delay in minutes between original CGM signal and predictions from ARIMA and LSTM models for different prediction horizons	30
4.1	Demographic and clinical characteristics ($Mean \pm SD$) of the subjects in the two clinical datasets.	40
4.2	Description of the Daily Physical Activity Features.	47
4.3	List of the Most Relevant Features Derived From the Sequential Forward Selection Algorithm.	47

4.4	The ROC-AUC Score and Sensitivity and Specificity of the Two Machine Learning Classifiers Evaluated During a 10-Fold Cross Validation in the Training and the Holdout Set.	49
4.5	Evaluation of the SVM algorithm in a new unseen dataset.	51
5.1	In-silico study results of bedtime intervention for nocturnal hypoglycaemia prevention. Statistically significant difference with respect to ‘No intervention’ ($P<0.05^*$, $P<0.01^{**}$). Statistically significant difference with respect to ‘Heuristic’ ($P<0.05^\#$, $P<0.01^{\#\#}$)	62
A.1	Full list of glucose-related features for NH prediction model (continued).	83
A.2	Full list of glucose-related features for NH prediction model (continued).	84
A.3	Diabetes related features for NH prediction model.	85

List of Figures

1.1	Visual representation of the contents	4
3.1	Illustration of an LSTM cell.	25
3.2	Blood glucose prediction performance of LSTM and ARIMA models on the 12 subjects of the OhioT1DM testing set. The lower and upper hinges of each boxplot represent the first (Q1) and third quartiles (Q3), respectively. The median is indicated by the central lines, and the whiskers extend to cover 1.5 times the interquartile range.	30
3.3	Blood glucose prediction performance of LSTM and ARIMA models on the subjects of the GVAS testing set. The lower and upper hinges of each boxplot represent the first (Q1) and third quartiles (Q3), respectively. The median is indicated by the central lines, and the whiskers extend to cover 1.5 times the interquartile range.	31
3.4	24-hour trajectories of glucose predictions over the 30 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.	32

3.5	24-hour trajectories of glucose predictions over the 60 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.	32
3.6	24-hour trajectories of glucose predictions over the 90 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.	33
3.7	24-hour trajectories of glucose predictions over the 120 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.	33
3.8	Blood glucose prediction performance of the LSTM model on the subjects of the GVAS testing set with diabetes (DM) and without diabetes (no-DM). The lower and upper hinges of each boxplot represent the first (Q1) and third quartiles (Q3), respectively. The median is indicated by the central lines, and the whiskers extend to cover 1.5 times the interquartile range.	34
3.9	Clark error grid (CEG) analysis on the GVAS dataset, showing the comparison between CGM measurements (x-axis) and glucose prediction values (y-axis) for (a) people with diabetes (DM) and (b) people without diabetes (no-DM).	35
4.1	SVMs try to find the optimal hyperplane which maximizes the margin	42
4.2	Colored boxes represent temporal windows used to extract daytime features (eg, blue: one hour prior to sleep, green: two hours prior to sleep, and orange: 12 hours prior to sleep). The gray shadowed area denotes nighttime.	46

4.3	Model development and internal validation workflow diagram..	48
4.4	ROC curve from the evaluation of the SVM binary classifier, using a 10-fold cross validation in the training set.	50
4.5	ROC curve from the evaluation of the SVM binary classifier on the holdout set.	50
4.6	ROC curve from the evaluation of the SVM binary classifier on the flash glucose monitoring data.	52
4.7	Recall-Precision curve from the evaluation of the SVM binary classifier on the flash glucose monitoring data.	52
5.1	Bar chart of the time spent in different blood glucose ranges for each one of the 10 adult participants during a period of 24 hours.	57
5.2	Bar chart of the different risk indexes for each one of the 10 adult participants during a period of 24 hours.	58
5.3	CVGA of the 10 adult participants during a period of 24 hours.	58
5.4	Participant recruitment and randomization for the IHART study.	66
5.5	Consort flow diagram of the ABC4D study.	68
5.6	Design of a Decision Support System for Nocturnal Hypoglycaemia Prevention.	70
5.7	Clinical trial design for a NH prevention study.	72

A.1	Visualization of data partitioning for individual and population models. Individual models were trained on the individual training sets and tested on the individual testing sets of the OhioT1DM dataset. For population-level models, we conducted an 80/20 split on the total training dataset, resulting in a training set and a separate validation set. The population models were trained using the actual training sets. Then the validation set was used for feature selection and hyperparameter tuning. Finally, the testing set was used to provide unbiased evaluation and the prediction results reported in this work. Population models were also tested in the GVAS dataset of hospitalized patients without retraining.	82
A.2	Visualization of data partitioning strategy utilized in this study for the development of NH prediction models. Initially, we conducted an 80/20 split on the ABC4D dataset, resulting in a training set and a separate hold-out testing set. Following this, a 10-fold cross-validation was performed to facilitate model validation and hyperparameter optimization. To ensure unbiased evaluation, the external IHART dataset was also incorporated.	85

Chapter 1

Introduction

1.1 Motivation

We live in an era where artificial intelligence (AI), defined by emeritus Stanford Professor John McCarthy in 1955 as ‘the science and engineering of making intelligent machines’, is disrupting and reshaping nearly every field and industry. From finance to entertainment, manufacturing, and transportation, AI is pushing the limits of what was previously considered impossible. Healthcare too, is another sector which has seen innovative breakthroughs and is expected to have a profound transformation driven by AI algorithms [1]. The ageing population, the growing burden of chronic diseases, along with staff shortages and the rising costs, are just some of the challenges that healthcare systems are facing today [2, 3]. Even though we are still in the early days of the adoption of AI in healthcare systems, there is growing evidence that AI applications can tackle effectively many tasks and help clinicians and patients alike [4].

One of the most successful applications of AI is in the field of computer vision. In one study, a convolutional neural network (CNN) trained in more than hundred thousand chest X-ray scans was able to detect pneumonia with an area under the curve (AUC) score of 76%. When compared with the results from radiologists who also examined the scans, it proved to be that the algorithm outperformed the doctors’ diagnosis [5]. A similar study was conducted in dermatology where an algorithm was trained in 130,000 patients’ images aiming to classify

skin cancer. The algorithm achieved an AUC around 95% both for carcinoma and melanoma detection. When compared with the accuracy of certified dermatologists it proved to be that the AI algorithm at least matched doctors' performance [6]. The list goes on with applications in ophthalmology [7], cardiology [8] and other areas. However, a notable example is a study which emphasized the synergy of clinicians and AI algorithms, as opposed to human versus machine. In this study, the task was to identify breast cancer, and results from the combination of pathologists and AI were more accurate than any of the two individual predictions alone[9].

ChatGPT (OpenAI, San Francisco, CA, USA) and many other similar generative artificial intelligence chatbots is another area which has a great potential to reshape healthcare. These chatbots are essentially deep learning algorithms that use the Transformer architecture [10], also called Large Language Models (LLM), which are trained on a huge amount of text data from the Internet, with the aim of generating text, based on user request. Such tools can be used as virtual assistants, helping clinicians to provide a diagnosis based on the symptoms of the patients and recommend appropriate interventions related to their condition [11]. One of the first of its kind randomized clinical trial aiming to evaluate the effectiveness of LLMs in enhancing physicians' diagnostic reasoning compared to traditional resources, showed that there were no significant differences between the two groups [12]. However, this study was limited by the relatively small number of participant physicians and by the fact that none of them were trained in prompt engineering, a concept which significantly improves the accuracy of results generated by LLMs.

In addition to AI tools designed for clinicians, there are also AI applications that help patients manage their health more effectively. A great example is diabetes, a very complicated metabolic disorder, which is characterized by glucose dysregulation. Medical devices such as continuous glucose monitoring (CGM) sensors, insulin pens and pumps have been a revolution in diabetes management and have become a necessity for many people. Harvesting the vast amount of data generated by these medical devices and combining them with data collected from health and wellness wearables and food intake apps [13], offer a unique opportunity to develop sophisticated solutions, such as intelligent DSS for diabetes management. DSS empowered by artificial intelligence algorithms can make predictions for future blood glucose levels, predict

adverse events before they occur (such as nocturnal hypoglycaemia), as well as recommend insulin doses and other interventions [14, 15]. Hence, they can add significant value in the management of diabetes and improve people's quality of life.

1.2 Research Objectives

The main research objectives of this thesis are summarized below:

- **Blood glucose level prediction:** The key research question here is to test how accurately we can predict future blood glucose levels in hospitalized patients, with or without diabetes. Extended work in this field has been done in outpatient populations, however the challenge with inpatients is that there are not a lot of available datasets and also, usually people stay in the hospital for a limited amount of time resulting in only a few data points per patient. The approach we follow is to use some of the best performing algorithms which have been proven to be effective in outpatients, and apply them in a proprietary clinical dataset of inpatients (Imperial College London, London, UK) to predict future glucose levels. Also whereas in most of the previous work the prediction horizon is up to 1 hour, here to try to extend this further to 90 and 120 minutes in the future.
- **Prediction of nocturnal hypoglycaemia (NH):** Here the objective is to develop and externally validate binary classifiers in order to predict before bedtime the occurrence of hypoglycaemia during sleep. The group of people we are most interested in, are people using multiple daily injection (MDI) insulin regimen, since this is the cohort which needs the most support in dealing with NH. We aim to find ways to best capture as many patterns as possible from glucose signals, and for this reason we use a time series feature engineering library to extract features across the statistical, temporal and spectral domains. Additionally, we aim to test if these models can be used in glucose sensors from different manufacturers and different sampling frequencies.
- **Nocturnal hypoglycaemia (NH) prevention:** The final objective has to do with the application of the previously developed models. The aim is to conduct an in-silico study

to test and compare prevention strategies for NH. To do so we use 1 year of data from 10 adult virtual subjects generated by a diabetes simulator. In this population there is an average of 8% time below range (TBR) overnight, and the idea is to try different prediction strategies, introduce 20 or 30 grams of rescue carbohydrate snacks (if needed) as intervention and measure the results. We also present briefly the architecture of a decision support system and the design of a real-world clinical trial.

1.3 Thesis Structure

Figure 1.1 shows a visual representation of the chapters in this thesis along with the recommended flow. The reader is introduced to the motivation and the research objectives in Chapter 1, before moving to Chapter 2 for a brief on diabetes mellitus and the challenges of this condition. In the following, the fundamental notions of machine learning and a literature review for blood glucose level prediction and nocturnal hypoglycaemia prediction are presented.

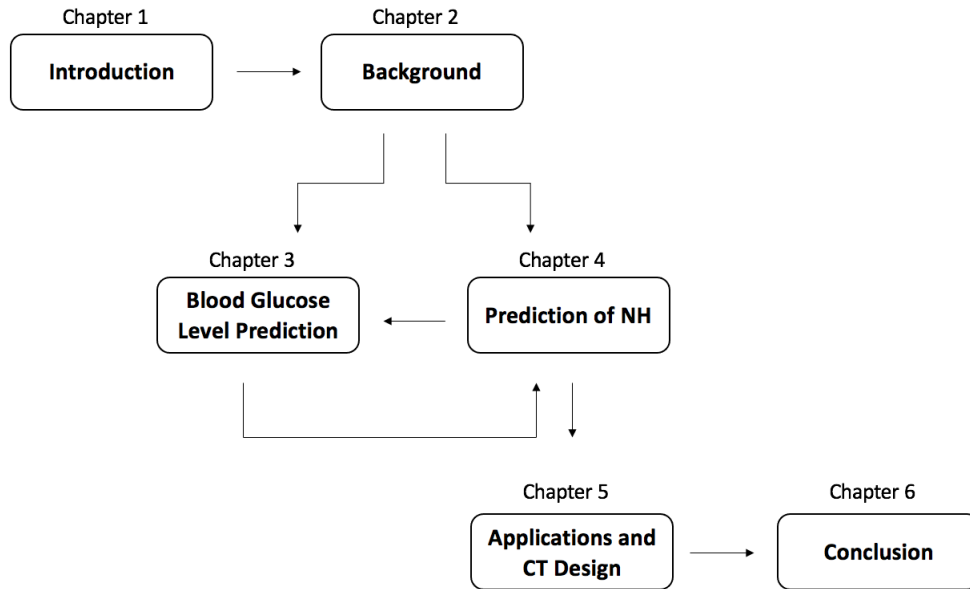


Figure 1.1: Visual representation of the contents

Each of the following three chapters refers to one research objective as discussed above. Chapter 3 and 4 are standalone in a sense that the reader can choose in which order to read each of

them. However, Chapter 4 is a prerequisite for Chapter 5. The reader needs to familiarize themselves with the binary classification of nocturnal hypoglycaemia, in order to understand the in-silico study and the prevention strategies in the form of bedtime carbohydrate snacks. Finally, Chapter 6 briefly summarizes the key results and the achievements of this work, and suggests opportunities for further research.

1.4 Thesis Contributions

An explicit list of the thesis contributions is presented below separately for each one of the three technical chapters:

- Chapter 3: We demonstrate that LSTM models trained on the Ohio T1DM dataset (a publicly available dataset containing continuous glucose monitoring (CGM) data from 12 adult outpatients over an 8-week period) can effectively predict blood glucose levels in hospitalized patients. Notably, these inpatients include individuals with type 2 diabetes (T2DM) as well as those without any history of diabetes. The hospitalized cohort consists of patients who were admitted following a confirmed stroke, with an average hospital stay of approximately three days. Our findings suggest that models trained on outpatient data can generalize well to inpatient settings, despite differences in health conditions, comorbidities, and care environments.
- Chapter 4: We propose a predictive model designed for nocturnal hypoglycaemia detection before bedtime, specifically targeting individuals with type 1 diabetes (T1DM) who follow a multiple daily injection (MDI) insulin regimen. The model is built exclusively on historical continuous glucose monitoring (CGM) data, eliminating the need for additional inputs such as insulin dosage, meal intake, or physical activity. This data-efficient approach makes it particularly well-suited for seamless integration into existing CGM or flash glucose monitoring systems without requiring external data sources. Furthermore, the model has undergone external validation, demonstrating its robustness and ability to generalize effectively to new populations. It also exhibits flexibility across different CGM

sampling frequencies and maintains high predictive accuracy when tested on glucose monitoring devices from various manufacturers. These characteristics enhance its practical applicability, making it a reliable tool for real-world nocturnal hypoglycaemia prediction and management.

- Chapter 5: In an in-silico study aimed at preventing nocturnal hypoglycaemia (NH), we compared the performance of our proposed predictive model against a simple heuristic algorithm that detects NH based on a predefined glucose threshold. While both approaches effectively reduced the time spent in the hypoglycaemic range (TBR), our model demonstrated superior performance by achieving a significantly greater reduction in TBR. Moreover, unlike the heuristic method, our model not only minimized hypoglycaemia but also led to an increase in the time in range (TIR), reflecting overall improved glucose control. Importantly, this improvement was achieved without any increase in time above range (TAR), meaning that our model did not elevate the risk of hyperglycaemia. This balance between reducing hypoglycaemia and maintaining overall glycaemic stability is a crucial clinical outcome, reinforcing the potential of our model for safe and effective nocturnal hypoglycaemia prevention.

Chapter 2

Background

2.1 Diabetes Mellitus

Diabetes mellitus is a non-communicable disease, ranked by the World Health Organization (WHO) as one of the top ten leading causes of death globally [16]. According to International Diabetes Federation (IDF), approximately 537 million adults (20-79 years) are currently living with diabetes in the world [17]. This is around 10.5% of the world's adult population. Diabetes is one of the fastest growing health challenges of the 21st century. The total number of people with diabetes is expected to reach 643 million by 2030 and 783 million by 2045, a dramatic 46% increase [17]. Diabetes is a growing global burden for individuals, families, and countries alike. It is estimated that a staggering amount of 966 billion dollars (USD) was spent on diabetes in 2021, which is about 10% of the global health expenditure [17, 18].

Diabetes mellitus is a chronic and complex metabolic disorder, characterised by abnormal glucose metabolism. It occurs when the pancreas is no longer able to produce insulin, or when the body cannot make sufficient use of the insulin it produces, a phenomenon known as insulin resistance. Insulin is a hormone which plays a central role to glucose homeostasis. It acts as a key to let glucose originated from the food we eat, pass from the blood stream into the cells in the body to produce energy. Not being able to produce insulin or use it effectively leads to raised glucose levels in the blood (known as hyperglycaemia). In the short-term, high glucose

levels cause dehydration, increased thirst and frequent urination. It is associated with fatigue, weakness and blurred vision, and increase the risk of infections and acute complications such as diabetic ketoacidosis. Over the long-term, high glucose levels are associated with damage to the body and failure of various organs and tissues, such as the heart, blood vessels, eyes, kidneys, nerves, teeth and gums.

Diabetes has a rich historical legacy dating back to ancient civilizations [19]. Greeks used the term ‘Diabetes’ or ‘Siphon’, referencing the pronounced polyuria characteristic of the condition. In India and Egypt individuals have experienced a sweet nature of urine, hence the Latin word ‘mellitus’. These historical notations signify the enduring prominence of diabetes as a significant medical concern across millennia.

Diabetes mellitus manifests in various forms, with type 1 and type 2 diabetes being the most recognized categories. Type 1 diabetes typically occurs early in life and results from autoimmune destruction of pancreatic β -cells, leading to an absolute deficit of insulin. In contrast, type 2 diabetes, predominantly seen in adults, arises due to impaired insulin function, where the cells do not react to insulin produced by the pancreas. Over 90% of people with diabetes have type 2 diabetes, which is driven by socioeconomic, demographic, environmental and genetic factors, such as urbanisation, the ageing population, decreasing levels of physical activity and increasing overweight and obesity prevalence [17]. Beyond type 1 and type 2 diabetes, the disorder encompasses an array of over 50 distinct variants. Among these, gestational diabetes, a temporary hyperglycaemic state during pregnancy, is a notable example.

The majority of people with diabetes live in Middle East and North Africa, Southeast Asia and the Western Pacific regions, in contrast with the lower levels observed in the African and European regions. A staggering three-quarters of people with diabetes reside in low- and middle-income countries, often with limited access to treatment [17, 20]. Notably, Europe has witnessed a decline in diabetes prevalence compared to historical levels, which researchers attribute to concerted efforts in lifestyle modification, fiscal policies such as tobacco taxation, and reduced advertising of unhealthy foods.

Diabetes mellitus is one of the most pressing non-communicable diseases globally and remains

a major medical challenge. An appreciation of its diverse forms, etiological factors, and the pivotal role of insulin in glycemic regulation is important for designing therapeutic interventions and evidence-based strategies. A multifaceted approach to prevention, management and personalised medicine is essential in order to enhance people's care and improve their quality of life.

2.2 Challenges of Diabetes Management

Managing diabetes can be challenging due to the complex nature of the condition. It requires a comprehensive approach that often involves a combination of medication, lifestyle changes, support from healthcare professionals, and ongoing self-management and education. Finding the right approach for each individual's unique needs and preferences can be daunting as it is not a one size fits all answer, due to the high inter-subject and intra-subject variability of the disease. Below we discuss the major tasks that people with diabetes have to undergo in order to manage their condition effectively. The primary focus is on type 1 diabetes management, however many of these also apply to type 2 diabetes cases as well.

Despite active research, type 1 diabetes cannot be prevented and currently there is no cure for it, so people rely on exogenous insulin for their survival. In order for people to maintain their blood glucose levels within a safe range (70-180mg/dL or 3.9-10mmol/L) [21], blood glucose monitoring is crucial. This often requires multiple daily fingerstick tests or continuous glucose monitoring (CGM). CGM sensors are minimally-invasive medical devices which are attached to the body and automatically measure subcutaneous glucose concentration converted into blood glucose levels with a frequency ranging from 1 up to 15 minutes, depending on the device [22]. They have seen a great adoption the recent years as they have facilitated diabetes management by moving away from the measurement of glycated hemoglobin (HbA1c) which has been the traditional method for assessing glycemic control [23].

Knowing their current glucose levels, the majority of people with type 1 diabetes inject insulin multiple times a day due to simplicity and cost effectiveness. In the UK and most European

countries MDI is the recommended therapy for most of the patients. Most of the remaining people (20% - 30% depending on the country) use insulin pumps which automatically infuse insulin into their body [24]. Continuous subcutaneous insulin infusion (CSII) pumps can be combined with CGM sensors as sensor-augmented therapy, or hybrid closed-loop systems [25, 26]. Hybrid closed-loop systems, also known as the artificial pancreas [27], even though they are commercially available, they have not been adopted widely (around 5% or 10% of the T1D population depending on the country), mainly due to economic and other human factors [28].

Balancing the timing, dosage, and potential side effects poses a daily challenge. Moreover, it is crucial to sustain a nutritious diet that regulates carbohydrate intake, monitors portion sizes, and controls calories consumption for effective diabetes management. Consistently making healthy dietary decisions can be particularly challenging, especially in social or stressful scenarios. Adding to the complexity, regular physical activity is essential for maintaining blood glucose levels. Nevertheless, finding the time and motivation to engage in exercise can be quite demanding.

In addition to modifying one's diet and participating in physical activity, sustaining lasting lifestyle adjustments plays a crucial role in diabetes management. This includes minimizing alcohol intake, quitting smoking, and achieving weight loss. Other essential steps involve actively seeking diabetes education and ensuring access to affordable tools and testing supplies, which may not be universally accessible, especially in areas with limited healthcare resources.

It is estimated that only 1 in 4 people achieve the recommended glycemic target [29]. Failing to achieve the glycemic target leads to hypoglycaemia or hyperglycaemia which over time can have microvascular complications, such as retinopathy, neuropathy, and nephropathy, alongside macrovascular afflictions such as cardiovascular disease. These complications, if left untreated, can culminate in blindness, cardiovascular events, neuropathic pain, kidney failure, and even premature mortality [30].

2.3 Fundamentals of Machine Learning

According to Arthur Samuel (1959), machine learning is defined as “the field of study that gives computers the ability to learn without being explicitly programmed”. The two main categories of machine learning are supervised and unsupervised machine learning. A brief introduction to each of these categories is presented in the following sections.

2.3.1 Supervised Machine Learning

Supervised machine learning, often called supervised learning, involves algorithms that learn to map inputs (x) to outputs (y). The defining feature of supervised learning is that the algorithm is trained on examples that include both the input and the correct output, or label (y). By repeatedly seeing correct pairs of inputs and their corresponding outputs, the algorithm learns to predict the output based on the input alone, without needing to be provided with the correct label during future predictions. So, once the model has learned from these input-output pairs (x and y), it can then take a completely new input (x^*) (never seen before) and attempt to generate the corresponding output (y^*).

There are two major types of supervised learning named regression and classification. Regression is a type of supervised learning that learns to predict numerical values from an infinite range of possibilities. Classification on the other hand, predicts a limited set of possible outcomes or categories, such as 0 or 1, rather than any numerical value from an infinite range. Classification algorithms are designed to predict categories, which may be non-numeric as well (e.g., identifying whether an image is of a cat or dog).

In the following chapters we present the development and validation of supervised Machine Learning algorithms, as well as statistical models, both for regression and classification. Specifically, for the task of blood glucose level prediction, we employ two very popular models for time series prediction, the Auto-Regressive Integrated Moving Average (ARIMA) and Recurrent Neural Networks with Long Short Term Memory (RNN LSTM). As shown in the next chapter, both models perform quite well especially for short-term predictive horizons, however

LSTM repeatedly outperforms ARIMA. This is due to LSTM's ability to handle non-linear relationships in the data (as opposed to ARIMA models which work best for linear time series), their ability to capture long-term dependencies using the memory cells and gates (more details on that in the next chapter), and their adaptability in different data distributions (as opposed to the ARIMA models which assume the time series are stationary).

The choice of an appropriate algorithm for any task depends heavily on the nature of the data and the transformations required to address each research question. While all the datasets used in the following chapters are time series (e.g., blood glucose levels over time, insulin units administered, and carbohydrate intake at specific times), they are processed in different ways depending on the task. For instance, in blood glucose level prediction, every data point in the time series serves as a potential output, giving the opportunity to data-hungry models to learn different patterns and relationships in order to make accurate predictions.

In contrast, nocturnal hypoglycemia prediction focuses on forecasting the overall trend of blood glucose levels throughout the night, rather than individual data points. This fundamental difference reshapes the data structure and significantly reduces its volume. As a result, in the first case, the dataset remains large enough to train and validate deep learning models, such as neural networks. However, in the latter case, the reduced dataset size makes traditional machine learning algorithms, such as Random Forests and Support Vector Machines, which are capable of learning patterns from smaller datasets, more suitable for the task.

2.3.2 Unsupervised Machine Learning

In supervised learning, the data includes both input values (x) and output labels (y). In contrast, unsupervised learning involves data that contains only input values (x) without any output labels (y), requiring the algorithm to discover structures or interesting features within the data. It is termed "unsupervised" because we do not guide the algorithm by supplying correct answers for each input. Instead, we allow the algorithm to explore the data independently to identify interesting patterns. For instance, an unsupervised learning algorithm may determine that the data can be divided into two or more distinct groups or clusters. Other

applications of unsupervised learning include anomaly detection, which is the identification of rare and unusual events; or dimensionality reduction, which allows us to reduce the size of a dataset without losing any important information.

2.4 Review of Blood Glucose Level Prediction

One of the fundamental tasks that people with diabetes have to undergo is to keep their glucose levels within the recommended euglycaemic range and avoid undesired events such as hypoglycaemia and hyperglycaemia. Having a good estimate of the future glucose values is crucial for taking the necessary actions to prevent these events. Hence, blood glucose level prediction is a research area which has attracted a lot of interest during the recent years. This has been enormously facilitated by the extensive CGM usage and the vast amount of data generated by these devices.

One of the earliest studies in the field is the work conducted by Sparacino et al [31]. Having as a main goal to prevent hypoglycaemic and hyperglycaemic events, they showed that glucose levels can be predicted for a short-term prediction horizon of up to 30 minutes. They used simple regression methods such as polynomial regression and autoregressive (AR) models based on historical blood glucose levels, and did not include information about insulin and carbohydrate intake, mainly due to the lack of data availability.

Later studies introduced physiological models [32] and more complex machine learning models [33, 34], as well as a combination of both [35], to predict future blood glucose levels. They also started including features related to insulin, meals and physical activity [36, 37] in an attempt to improve models performance. On the one hand, some researchers have reported an increase in models' performance, while others claim that these data points often add noise to the models mainly due to human error and scarcity of information [38, 39].

Due to the massive success of deep learning in various fields, artificial neural networks are another class of very powerful methods that has attracted a lot of attention in the field of blood glucose levels prediction [40]. In one of the first studies, Mougiakakou et al [41] used

compartmental models to extract features related to glucose, insulin and meal information and fed them into feed-forward and recurrent neural networks. Even though the model performed well for glucose values greater than 70 mg/dL, it underperformed below the target range. The authors attributed this to the lack of training examples.

In more recent studies various architectures of neural networks have been tested aiming to make accurate predictions of future blood glucose levels [40]. A noteworthy example is GluNet, a dilated convolutional neural network (CNN) trained on historical glucose measurements, insulin doses, carbohydrate intake and other important factors [42]. The authors reported an RMSE value of 19.28 mg/dL for 30 minutes prediction horizon and 31.83 mg/dL for 60 minutes prediction horizon when tested in two clinical datasets. They also showed that the model outperforms more traditional methods such as support vector regression (SVR), latent variable with exogenous input (LVX) and auto regression with exogenous inputs (ARX).

Other studies have also attempted a contrast between different models [43, 44, 45, 46, 47]. Despite the fact that a comparison among them is not fair, since most works have used different clinical datasets for model development and evaluation, a common denominator is the dominance of the recurrent neural networks (RNN) with long short-term memory (LSTM) architecture. Their adaptability to varying sequence lengths, flexibility in handling different time series characteristics, and ability to learn temporal patterns contribute to their success in many applications involving sequential data.

Ref	Model	RMSE (30mins)	RMSE (60mins)
J. Chen et al [48]	DRNN	19.04mg/dL	-
J. Martinsson et al [49]	LSTM	18.86mg/dL	31.40mg/dL
T. Yang et al [50]	MS-LSTM	19.04mg/dL	32.02mg/dL
C. Midroni et al [51]	LSTM (and XGBoost)	23.48mg/dL	-
R. Bevan et al [52]	LSTM	18.23mg/dL	31.10mg/dL
M. Mayo et al [53]	DNN with 12 LSTM units	19.80mg/dL	34.00mg/dL
G. Cappon et al [54]	bidirectional LSTM	20.20mg/dL	34.19mg/dL

Table 2.1: Summary of Previous Studies from the KDD BGLP Challenge using variations of LSTM models along with performance in 30 and 60 minutes prediction horizon.

The outstanding performance of LSTM models has also been observed in two blood glucose level prediction challenges organised as part of the International Workshop on Knowledge Discovery

in Healthcare Data (KDD) in 2018 and 2020 [55, 56]. In both challenges participants were given the OhioT1DM dataset [57] containing data from 12 participants (6 each year) and were asked to develop and test their models in the predefined training and testing sets respectively. The final evaluation was performed based on the root mean square error (RMSE) for the 30 minutes and 60 minutes prediction horizon. Some of the top performing models were LSTMs or variations of those [48, 58, 49, 50, 51, 52, 53, 54]. The performance of these models in the 30 and 60 minutes prediction horizon is summarized in Table 2.1.

Last but not least a recent study by Herrero et al [15] showcased the Accu-Chek SmartGuide Predict App (Roche Holding AG, Basel, Switzerland). Among other functionalities this innovative digital companion app offers a 2 hours real-time glucose level prediction along with uncertainty estimates aiming to empower users to better manage their diabetes. The model used is a gated recurrent unit (GRU) based neural network, which is also an RNN with similar but less complex architecture than the LSTM's. Using features from CGM data, insulin, meals and the time of the day they evaluated the model in different datasets achieving a remarkable average of 96.3% of data points in zones A and B of the Consensus Error Grid (accurate or clinically acceptable results). Their results were even better for shorter prediction horizons of 30, 60, and 90 minutes with 99.8%, 99.3%, and 98.7% respectively.

2.5 Review of Nocturnal Hypoglycaemia Prediction

One of the most common approaches for insulin administration, necessary for diabetes management, is multiple daily injections (MDI). Within this framework, a substantial hurdle arises in the form of hypoglycaemia [23], a condition defined by blood glucose levels dropping below the threshold of 70 mg/dL. Notably, individuals with type 1 diabetes (T1D) often encounter hypoglycemic episodes during the nocturnal period [59], despite the availability of sophisticated sensing technologies [60] designed to detect such adverse events. This phenomenon, referred to as nocturnal hypoglycemia (NH), poses a considerable challenge [61] as awareness of hypoglycaemia is significantly diminished during sleep, and prolonged exposure to NH may lead to

impaired awareness, and even to the ‘dead-in-bed’ syndrome.

The widespread integration of continuous glucose monitors (CGMs) has given rise to extensive datasets, presenting an opportunity for the application of supervised machine learning (ML) algorithms in predicting adverse events before they occur. While prior research predominantly focused on short-term NH prediction [62, 63, 64], a notable gap exists in literature proposing models for forecasting NH over more extended periods, such as during the night time. A summary of such studies is presented in Table 2.2.

Ref	Population	Nights	Algorithms	ROC-AUC
Vu et al [65]	>10,000	>1,000,000	RF	84%
Jensen et al [66]	463	-	LDA	79%
Bertachi et al [67]	10	84	MLP,SVM	>70%
Güemes et al [68]	6	-	RF,NN,SVM,ETC	70%
Mosquera-Lopez et al [69]	124	23,000	SVR	77%
Afentakis et al [70]	37	6,000	RF,SVM	79%
Herrero et al [15]	221	>13,000	XGBoost	86%

Table 2.2: Summary of Previous Studies on Nocturnal Hypoglycaemia Prediction.

In a noteworthy study, Vu et al [65] harnessed a substantial dataset comprising of 10,000 CGM users and spanning over a million nights to construct a Random Forest (RF) binary classifier, demonstrating efficacy in predicting NH occurrences with a commendable ROC-AUC score of 84%. Similarly, Jensen et al [66] delved into a clinical trial dataset involving 463 people with T1D to predict level-2 NH (≤ 54 mg/dL). Their methodology involved extracting features from both daytime and the three consecutive days leading up to the night, training a linear discriminant function algorithm that achieved a noteworthy 79% ROC-AUC score at midnight.

Taking an individualized approach, Bertachi et al [67] developed prediction models tailored to detect over 70% of NH events in T1D individuals. Their algorithms, trained on a 12-week clinical study involving 10 participants, incorporated a rich set of 29 features derived from CGM signals, insulin dosages, meal intake data, and activity tracking. The support vector machine (SVM) algorithm emerged as the superior performer over the multilayer perceptron (MLP) in this specific population.

Güemes et al [68] proposed a method grounded in binary classifiers to predict the quality of

overnight glycemic control in individuals with T1D. Leveraging publicly available clinical data [57], they trained multiple ML models aimed at predicting nocturnal hypoglycaemia or hyperglycaemia, alongside quantifying the percentage of time spent within the optimal glycemic range (%TIR). Their comprehensive approach resulted in an overall ROC-AUC score of approximately 70%.

Drawing insights from a vast dataset of 124 individuals with T1D, Mosquera-Lopez et al [69] employed glucose, insulin, and meal information to craft a support vector regression (SVR) model. Through in silico testing, their results indicated a remarkable 77% reduction in NH events without compromising the time spent within the target glycemic range (TIR).

All the aforementioned studies are centered on people utilizing CGMs for glucose monitoring, and pumps for insulin infusion. Only in one study [67] focuses on participants on MDI therapy which is the most vulnerable population for NH. Additionally, external validation is only applied to the work of Mosquera Lopez et al [69] where two external datasets are utilised.

The two issues above are addressed in the study conducted by Afentakis et al [70]. The authors explored an open-source framework for feature extraction from time series data, and used it for blood glucose time series. By employing two distinct datasets from individuals utilizing both CGMs and MDI, the objective is to develop and externally validate ML models, ensuring robust generalization as well as applicability across sensors from diverse manufacturers. The proposed model achieves a ROC-AUC score of more than 79% and holds significant promise as a comprehensive bedtime decision support tool for individuals managing T1D, effectively minimizing the risk of nocturnal hypoglycemia.

Another study which addresses the lack of external validation as well as targeting the population with a MDI insulin regimen, is the work by Herrero et al [15]. As mentioned in the previous section, the authors present the Accu-Chek SmartGuide Predict App (Roche Holding AG, Basel, Switzerland) which also includes a night low predict functionality. The model predicts the risk of NH for the first (0-3.5 hours) and the second half (3.5-7 hours) of the night offering an estimate of the approximate time, along with recommendations for intervention. The model used is an XGBoost classifier developed in a dataset comprising 221 adults with T1D on MDI

therapy and externally validated in two other datasets including people with type 1 and type 2 diabetes. An outstanding performance of 85.9% (ROC AUC) was achieved. Even though a fair comparison between the studies is not possible, this is the highest ROC AUC score ever reported in the literature, in the task of NH prediction.

2.6 Review of Decision Support Systems in Diabetes Management

Over the last few years, the growing popularity of mobile health applications, combined with major advancements in smartphone technology, has resulted in a sharp increase in diabetes-related apps across iOS and Android marketplaces [71]. Despite the growing number of decision support apps for diabetes management, only a limited few have undergone clinical trials. Some incorporate basic algorithms to assist with bolus insulin calculations, easing the demands of intensive diabetes management. However, most function primarily as simple electronic diaries or data visualization tools, often lacking integration with wearable or medical devices, or broader diabetes management systems. A handful of apps leverage image processing technology to enhance carbohydrate estimation accuracy and improve postprandial blood glucose control [72]. A recent review highlights that while smartphone apps offer various advantages for diabetes management, their design and features should be more tailored to the needs of individuals with type 1 diabetes. Enhancing usability and functionality could lead to better clinical outcomes and an improved quality of life [73].

The Internet of Things (IoT) has revolutionized blood glucose monitoring and glycemic control, enhancing diabetes management through seamless integration of wearable devices [74, 75]. Advanced artificial pancreas (AP) systems now interconnect continuous glucose monitors (CGMs), insulin pumps, smart insulin pens, glucagon pumps, and physiological wristbands to track vital signals with greater accuracy and efficiency. Herrero et al. [76], introduced the Bio-inspired Artificial Pancreas, featuring a customized handheld device designed to run glycemic control algorithms on a microcontroller. This device communicates via Bluetooth with a continuous

glucose monitor (CGM), an insulin pump, and a dedicated smartphone application, which also offers cloud-based remote monitoring. The system's effectiveness was assessed using the UVA/Padova T1D simulator and subsequently validated through a clinical trial [77].

Last but not least, Accu-Chek SmartGuide Predict app [15], a very promising digital companion for people with T1DM offers three predictive functionalities to its users. Forecasting of future blood glucose levels over a two hours prediction horizon, a 30 minutes hypoglycaemia detection, as well as nocturnal hypoglycaemia prediction aiming to be used for bedtime interventions. Although the system has demonstrated state-of-the-art performance, a clinical trial has yet to be conducted.

2.7 Conclusion

In the beginning of this chapter, the epidemiology of diabetes mellitus was discussed along with the challenges that people with diabetes face in their daily lives. Later on, a literature review of blood glucose level prediction (BGLP) algorithms was presented, with a specific focus on two BGLP challenges based on a clinical dataset of outpatients. The outstanding performance of LSTMs was highlighted as one of the top performing algorithms for this task. In the following, a review of the previous work on nocturnal hypoglycaemia prediction was discussed, noting the significant gap in the literature, as opposed to short term hypoglycaemia or hyperglycaemia detection. A summary of the previous studies was also presented in a table along with some characteristics including the population volume, the algorithms used and the models' performance.

Chapter 3

Blood Glucose Level Prediction

3.1 Introduction

The topic of this chapter lies in the area of models development for future blood glucose levels prediction. This is a heavily researched area where lots of different approaches have been tried and tested in various datasets. Our primary goal in this chapter is not to compare the performance of multiple models, or propose a superior algorithm over the state-of-the-art. On the contrary, we use some of the best performing models from the literature, such as recurrent neural networks [78] with long short-term memory (LSTM) [79, 80], as well as autoregressive integrated moving average (ARIMA) [81] models (which are statistical models very effective for time series analysis), in order to test some clinical research hypotheses.

The vast majority of the previous work has been focused on outpatient populations mainly due to the limited availability of glucose data for hospitalised patients. However, in more recent years CGM usage has been adopted in many clinical scenarios and in hospital settings [82, 83, 84], especially during the COVID-19 pandemic [85, 86]. It has also been used by individuals without diabetes [87] in order to support dietary strategies and improve metabolic complications [88].

The research hypothesis we aim to answer is whether or not blood glucose level prediction

models, trained and tested in outpatients with T1DM, can be used to make predictions about future blood glucose levels of inpatients with T2DM or without diabetes at all. The rationale behind this hypothesis is that these models have shown great performance when the CGM signal is relatively smooth without abrupt changes and fluctuations. Since people without diabetes at all, or those with T2DM, are expected to have much better glucose regulation, as opposed to people with T1DM, we assume that the aforementioned models will be able to generalize well in this new unseen population.

In this chapter we present some evidence that we can predict future blood glucose levels of those in the hospital bed. Around one in four people in hospitals have diabetes, and proactive prevention of extremes of glucose is incredibly important for mortality, morbidity, cost and resource utilisation. For this reason we apply models used in outpatient populations, to a proprietary clinical dataset of inpatients (Imperial College London, London, UK), and show that these models can be very effective in estimating future blood glucose values. Additionally, we evaluate these models in people with and without diabetes and show that their performance does not deteriorate in neither of the two population groups.

Most of the previous studies on blood glucose level prediction has been focused on short-term (up to 1 hour) prediction horizon [40]. In this work, we extend the usual short-term predictive horizon by generating predictions for 90 minutes and 120 minutes into the future.

3.2 Data

3.2.1 OhioT1DM dataset

The first dataset used in this work is the OhioT1DM dataset which is a publicly available dataset first released in 2018 and updated in 2020 [57]. The dataset contains data from 12 adults with T1D over a period of eight weeks. Each participant in the study utilized a Medtronic Enlite Continuous Glucose Monitoring (CGM) system paired with either a Medtronic 530G or 630G insulin pump. The participants recorded their daily events using a dedicated smartphone

application. Additionally, a subset of individuals opted to wear either Basis Peak or Empatica Embrace bands to gather comprehensive vital signs data. The collected data has been split into training and testing sets, with the training set encompassing around 80% of the total samples and the testing set the remaining 20%. In this work we have kept this suggested division which ensures a robust evaluation of the models performance.

3.2.2 GVAS dataset

The second dataset (GVAS) originates from a previous observational study that sought to compare the mean absolute glucose (MAG) change within the initial 72 hours post a confirmed stroke in two distinct groups: individuals with and without diabetes [89]. The study encompassed 67 adult participants with clinically suspected ischemic stroke symptoms within a 12-hour window. Inclusion criteria considered known diabetes, HbA1c greater than 42, or absence of known diabetes. Exclusion criteria comprised hemorrhagic stroke, pregnancy, premorbid MRS greater than 2, age exceeding 80 years, and terminal conditions with a life expectancy of less than a year. The primary metric, MAG change, was assessed without any intervention. The study aimed to shed light on glucose dynamics during the early stages of stroke and potential variations between stroke patients with and without diabetes. The results indicated a correlation between glycemic variability and a decline in functional status upon discharge post-acute stroke. Data from patients with less than 20 hours of information were excluded, and any missing or duplicated records were eliminated, resulting in 45 subjects for analysis. The reason we chose not to impute any missing values but rather to keep only the existing data points is because we did not want to introduce external noise, since the volume of the available data was significant for the validation (less than 5% of missing glucose values).

In Table 3.1 we can see an overview of the demographics and clinical characteristics of the populations of the two datasets used in this work. Apart from the age, gender, insulin regimen and mean value of glucose levels of participants, the most common clinical CGM metrics have been calculated as recommended by the international consensus for applications in clinical diabetes care [21]. These include glucose management indicator (GMI) [90], time below

range (TBR), time in range (TIR), time above range (TAR), blood glucose index, and coefficient of variation (CV). The recently proposed GMI is a valuable alternative to the traditional laboratory-measured HbA1c, when CGM data are available.

Characteristics	OhioT1DM	GVAS
Age (years)	50 ± 10	74 ± 12
Gender (male/female)	7/5	27/18
Insulin regimen (CSII/MDI)	12/0	N/A
Mean glucose level (mg/dL)	162.3 ± 20.0	140.3 ± 48.2
GMI (%)	7.2 ± 0.5	6.7 ± 1.2
eA1c (mg/dL)	131.4 ± 12.6	117.0 ± 30.6
TBR (<54 mg/dL) (%)	0.4 ± 0.4	1.2 ± 3.8
TBR (54-69 mg/dL) (%)	2.3 ± 2.0	2.5 ± 10.5
TIR (70-180 mg/dL) (%)	62.9 ± 13.2	80.6 ± 28.5
TAR (181-250 mg/dL) (%)	25.7 ± 8.0	10.2 ± 15.0
TAR (>250 mg/dL) (%)	8.8 ± 8.5	5.5 ± 18.2
Low blood glucose index	0.7 ± 0.4	1.0 ± 2.3
High blood glucose index	7.4 ± 3.3	4.4 ± 8.0
Inter-day CV (%)	34.3 ± 4.6	18.8 ± 6.9
Intra-day CV (%)	29.2 ± 4.5	16.0 ± 6.9

Table 3.1: Demographic and clinical characteristics ($Mean \pm SD$) of the subjects in the two clinical datasets.

3.3 Algorithms and Model Development

In a regression model, the target value y_t at a given time t is predicted based on the features (independent variables) at that specific time t , independently of the values of y at other time points. The foundational statistical assumption in regression is that y_t can be expressed as the sum of a deterministic component and a stochastic error term. The stochastic terms are assumed to be independently added to the deterministic term for each y_t . However, this assumption is not suitable for time series data, where the value of y_t may be influenced by previous values such as y_{t-1} , y_{t-2} , y_{t-3} etc.

3.3.1 ARIMA(p,d,q)

ARIMA models, also known as Box-Jenkins models [91], are a class of models aimed at expressing y_t as a function of preceding y values and error terms, often involving differencing. An ARIMA(p, d, q) model is comprised of an Autoregressive (AR) component with order p ($AR(p)$), linking y_t to the last p y values, a Moving Average (MA) component with order q ($MA(q)$), connecting y_t to the previous q error terms, and an Integrated (I) component with order d ($I(d)$), representing d steps of differencing. Differencing is employed to make the time series stationary, as the other two components perform more effectively on a stationary time series. When combined, these three components define y_t as:

$$\hat{y}_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_{d+p} y_{t-d-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} + \mu \quad (3.1)$$

The hyper-parameters p , q , and d are commonly manually chosen by inspecting the data or experimenting with various values. This experimentation often involves a grid search over a set of predefined values. In this work in order to set the hyper-parameters p , q , and d , a cross validation in the training set of the OhioT1DM dataset was performed over the sets $P = \{1, 2, \dots, 6\}$, $D = \{0, 1, \dots, 6\}$ and $Q = \{0, 1, \dots, 6\}$. Alternative tools are the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) which can assist in selecting the AR and MA components.

Once the hyper-parameters are chosen, the model must acquire knowledge from the data to determine its parameters $(\phi_1, \phi_2, \dots, \phi_{d+p}, \theta_1, \theta_2, \dots, \theta_q)$. Once these parameters are learned, forecasting future values beyond time T involves the application of Equation 3.1. Specifically, starting with $y_1, y_2, \dots, y_T$ we can compute $\hat{y}_{T+1}$ using the equation, then proceed to compute $\hat{y}_{T+2}$ based on the obtained result, and so forth. It's important to note that as the prediction horizon T extends, the accuracy of our predictions may deteriorate, as forecasting $\hat{y}_{T+h}$ relies on previously predicted values from y_{T+1} to y_{T+h-1} , rather than the actual values.

3.3.2 RNN LSTM

As mentioned in Section 2.4, one category of models which has achieved state-of-the-art performance in blood glucose level prediction is the recurrent neural networks (RNN) with long short-term memory (LSTM) architecture [92, 93]. LSTMs address the vanishing gradient problem and are able to capture long-term dependencies in the data. An LSTM is actually an RNN which introduces a memory cell structure, including an input gate, a forget gate and an output gate, which control what information enters and exits the memory cell. An illustration of an LSTM memory cell is shown in Figure 3.1.

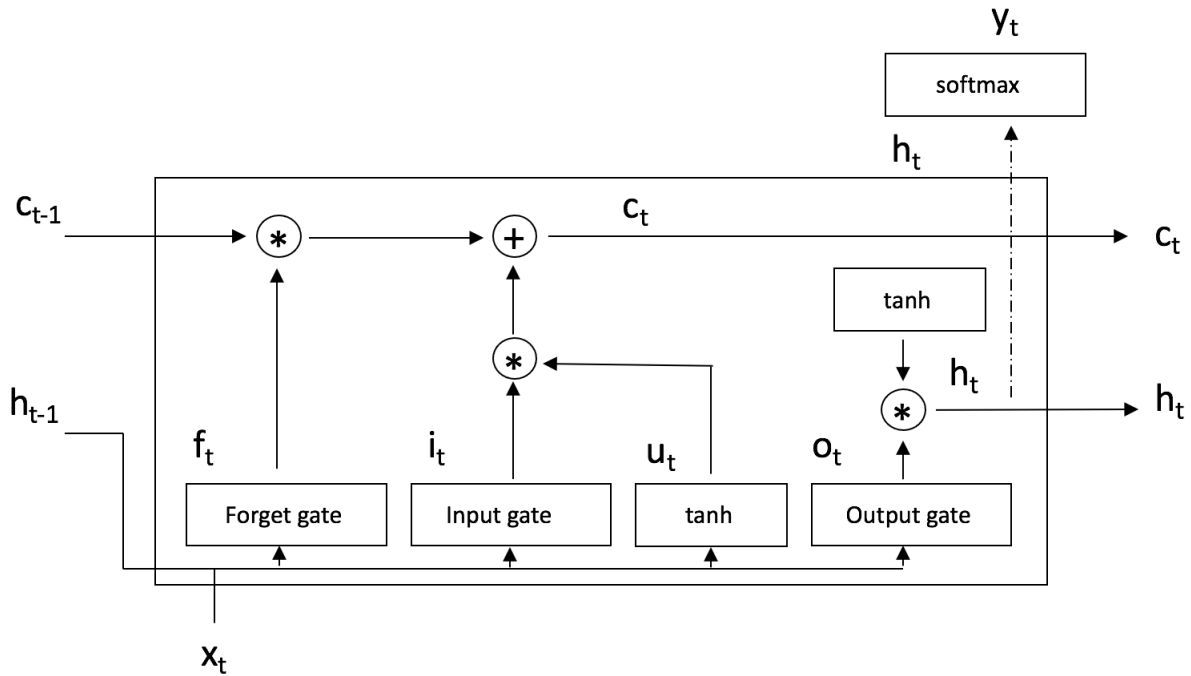


Figure 3.1: Illustration of an LSTM cell.

An LSTM not only can add or remove information from the cell state, but also can keep it unchanged from one time step to the other. Specifically, the forget gate is responsible for controlling which parts of the internal memory will be kept or not. The input gate selects the parts of the input that will be stored in the internal memory and the output gate determines those that will be passed on to the next neurons. Each of the three gates take values in the range $[0, 1]$, with zero and one representing closed and open gates respectively.

The following operations are applied at each step of the sequence in order to compute the new internal memory $\mathbf{c}_t$ and the cell output $\mathbf{h}_t$. In the below $\mathbf{x}_t$ is the input to the memory cell at time t and W_i, W_f, W_o, W_u as well as U_i, U_f, U_o, U_u are weight matrices for input and hidden state. Finally, $\mathbf{b}_i, \mathbf{b}_f, \mathbf{b}_o, \mathbf{b}_u$ are bias vectors.

$$\mathbf{i}_t = \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (3.2)$$

$$\mathbf{f}_t = \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (3.3)$$

$$\mathbf{o}_t = \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (3.4)$$

$$\mathbf{u}_t = \tanh(W_u \mathbf{x}_t + U_u \mathbf{h}_{t-1} + \mathbf{b}_u) \quad (3.5)$$

$$\mathbf{c}_t = \mathbf{i}_t * \mathbf{u}_t + \mathbf{f}_t * \mathbf{c}_{t-1} \quad (3.6)$$

$$\mathbf{h}_t = \mathbf{o}_t * \tanh(\mathbf{c}_t) \quad (3.7)$$

The model takes as input a sequence of glucose data, provided by the CGM sensor [49]. The concluding LSTM cell in the sequence produces an output vector, which is then fed into a fully connected neural network featuring two hidden dense layers and a single output layer. The hidden layers encompass 512 and 256 neurons, with rectified linear activations, and dropout rates of 20% and 30%, respectively. These dropout layers are introduced to avoid overfitting the training data.

The output layer is composed of two neurons, one with a linear activation and the other with an exponential activation. The linear activation is used for the mean and the exponential for the standard deviation of a univariate Gaussian distribution as shown in equations 3.8 and 3.9, where $\mathbf{h}_{fc}$ is the output of the last hidden layer. This way the output contains a confidence of the prediction for the glucose value after time T , where $T \in \{30, 60, 90, 120\}$ minutes.

$$\mu = W_1 \mathbf{h}_{fc} + \mathbf{b}_1 \quad (3.8)$$

$$\sigma = \exp(W_2 \mathbf{h}_{fc} + \mathbf{b}_2) \quad (3.9)$$

As loss function, the negative log-likelihood (NLL) function has been used, derived from the Gaussian distribution as shown in equation 3.10

$$L = \frac{1}{k} \sum_{i=0}^k -\log(N(y_i|\mu_i, \sigma_i^2)) \quad (3.10)$$

Both the individual and population LSTM and ARIMA models presented in the next section were trained and fine-tuned in the training set of the OhioT1DM dataset, and tested in the OhioT1DM testing set. Later, and the two population models (ARIMA(2,0,0) and LSTM) are tested in the GVAS dataset without further retraining. A visual representation of how the different datasets were used for the training and testing of different models is shown in Figure A.1 in the Appendix.

3.4 Results and Discussion

In Tables 3.2, 3.3, 3.4 and 3.5, we present the performance of the personalised ARIMA and LSTM models on the testing set of each one of the participants of the OhioT1DM dataset. The primary evaluation metric used is the root mean square error (RMSE). This is due to its widespread use in the literature [40] and its direct representation of prediction accuracy. The tables below contain results for the different prediction horizons, namely 30, 60, 90 and 120 minutes respectively. On average, between the two personalised models, LSTM consistently outperforms ARIMA with 19.48, 33.39, 45.41 and 49.40 mg/dL of the the former, compared with 26.58, 48.04, 59.90 and 80.29 mg/dL of the latter, for 30, 60, 90 and 120 minutes respectively. It is worth mentioning that the lowest RMSE value achieved among participants for the 30 minutes prediction is 16.15 mg/dL and 27.97 mg/dL for the 60 minutes prediction. In both cases results refer to participant 570, something that has been observed in other studies as well [49] and has to do with the intra-subject variability of glucose levels.

In Figure 3.2 we can see the performance of the population models (ARIMA and LSTM) on the same testing sets of the OhioT1DM dataset for all the 12 subjects and for all the different

OhioT1DM	ARIMA		LSTM
Participant id	orders (p,d,q)	RMSE (mg/dL)	RMSE (mg/dL)
559	3,2,0	26.76	19.94
563	2,0,0	23.49	18.44
570	4,0,4	24.05	16.15
575	5,0,4	22.63	22.66
588	1,0,3	26.73	19.01
591	2,0,4	30.51	21.91
540	2,0,0	30.19	21.36
544	4,0,3	24.34	17.36
552	4,1,4	23.26	16.46
567	3,2,0	30.61	20.9
584	2,0,0	28.67	22.11
596	2,0,0	27.78	17.46

Table 3.2: Personalised models performance with 30 minutes prediction horizon.

OhioT1DM	ARIMA		LSTM
Participant id	orders (p,d,q)	RMSE (mg/dL)	RMSE (mg/dL)
559	3,2,0	47.60	34.53
563	2,0,0	50.97	30.96
570	4,0,4	65.67	27.97
575	5,0,4	41.62	37.48
588	1,0,3	44.23	31.25
591	2,0,4	42.66	33.85
540	2,0,0	49.61	41.3
544	4,0,3	47.01	30.62
552	4,1,4	44.7	30.06
567	3,2,0	56.42	36.23
584	2,0,0	45.54	36.99
596	2,0,0	40.47	29.38

Table 3.3: Personalised models performance with 60 minutes prediction horizon.

prediction horizons. The black horizontal line of each boxplot indicates the median value of RMSE, whereas the lower and upper hinges represent the first (Q1) and third (Q3) quartiles respectively. Whiskers extend further to cover up to 1.5 times the interquartile range and observations out of this range are considered to be outliers.

Similar to the personalised models, in population models the LSTM outperforms the ARIMA model. However both models show little variability in the RMSE values for the 30 and 60 minutes prediction horizon. It is for the longer prediction horizons of 90 and 120 minutes where the predictions of the ARIMA model become more dispersed, which is also accompanied

OhioT1DM	ARIMA		LSTM
Participant id	orders (p,d,q)	RMSE (mg/dL)	RMSE (mg/dL)
559	3,2,0	60.74	46.76
563	2,0,0	78.46	41.26
570	4,0,4	60.89	35.4
575	5,0,4	50.09	46.33
588	1,0,3	57.8	39.66
591	2,0,4	51.11	42.84
540	2,0,0	66.06	84.34
544	4,0,3	58.81	39.78
552	4,1,4	59.38	37.86
567	3,2,0	71.09	45.35
584	2,0,0	52.63	47.13
596	2,0,0	51.72	38.16

Table 3.4: Personalised models performance with 90 minutes prediction horizon.

OhioT1DM	ARIMA		LSTM
Participant id	orders (p,d,q)	RMSE (mg/dL)	RMSE (mg/dL)
559	3,2,0	73.39	55.75
563	2,0,0	105.95	47.93
570	4,0,4	74.18	41.59
575	5,0,4	73.99	52.72
588	1,0,3	68.30	44.12
591	2,0,4	58.14	44.78
540	2,0,0	85.69	65.17
544	4,0,3	95.41	46.51
552	4,1,4	97.43	43.85
567	3,2,0	110.09	50.96
584	2,0,0	60.25	54.71
596	2,0,0	60.63	44.75

Table 3.5: Personalised models performance with 120 minutes prediction horizon.

with a skewed distribution of the error values.

In Figure 3.3 we show the performance of the two population models (ARIMA and LSTM) on the testing sets of the participants of the GVAS study. We can see that across all the different prediction horizons both models perform quite well when comparing the results with the OhioT1DM population models (Figure 3.2) as well as other similar studies of the Blood Glucose Level Prediction Challenge [57]. This can be further observed in Figures 3.4, 3.5, 3.6, and 3.7, with the 24-hour trajectories of glucose predictions over the different prediction horizons. The red line represents actual CGM measurements for a total of 288 data points (24

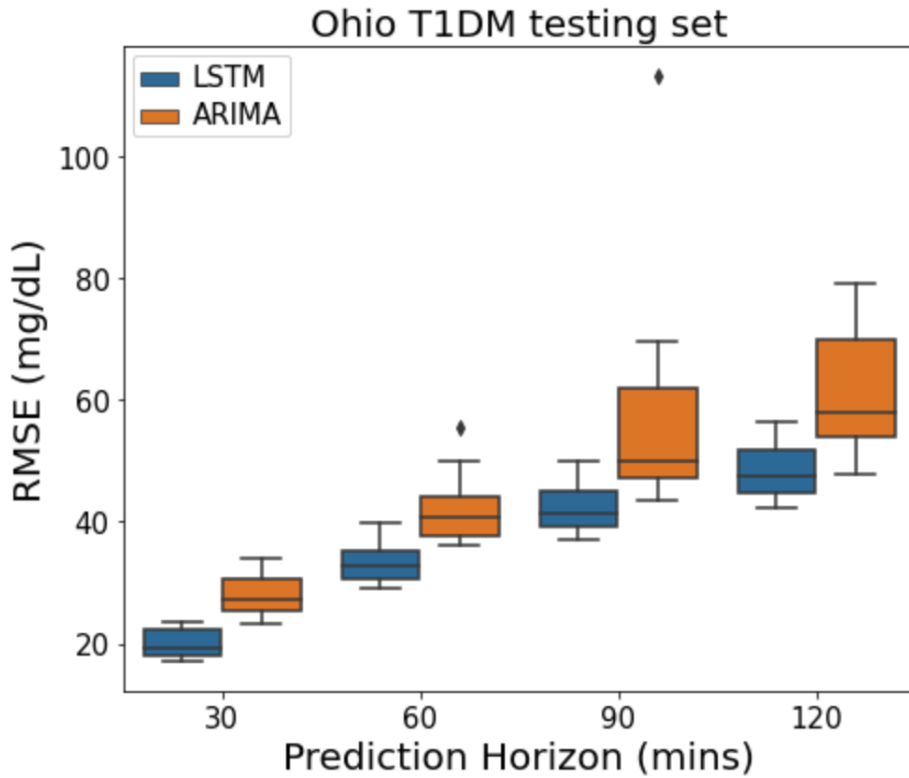


Figure 3.2: Blood glucose prediction performance of LSTM and ARIMA models on the 12 subjects of the OhioT1DM testing set. The lower and upper hinges of each boxplot represent the first (Q1) and third quartiles (Q3), respectively. The median is indicated by the central lines, and the whiskers extend to cover 1.5 times the interquartile range.

hours). The yellow line corresponds to the ARIMA predictions and the blue line along with the light-blue shaded area, to the mean and standard deviation of predictions generated by the LSTM model. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.

Additionally, as it can be observed in Figures 3.4, 3.5, 3.6, and 3.7 there is a time lag between the actual CGM signal and the predictions, which become more obvious for the longer prediction horizons. In Table 3.6 we have quantified this Prediction Time Delay (PTD) in minutes for both ARIMA and LSTM and for each one of the four prediction horizons.

	Prediction Time Delay (PTD) in minutes			
	30 mins	60 mins	90 mins	120 mins
ARIMA	10.8	18.9	25.2	35.1
LSTM	8.6	16.0	19.6	27.0

Table 3.6: Prediction Time Delay in minutes between original CGM signal and predictions from ARIMA and LSTM models for different prediction horizons

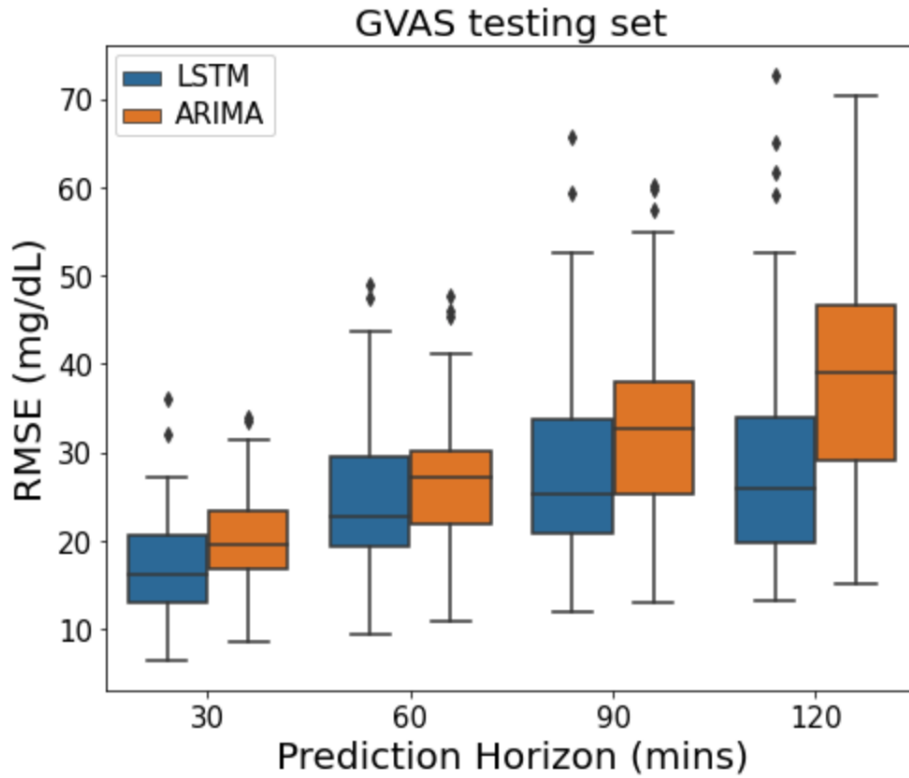


Figure 3.3: Blood glucose prediction performance of LSTM and ARIMA models on the subjects of the GVAS testing set. The lower and upper hinges of each boxplot represent the first (Q1) and third quartiles (Q3), respectively. The median is indicated by the central lines, and the whiskers extend to cover 1.5 times the interquartile range.

In this work we used the OhioT1DM dataset as a benchmark for models performance, however a head-to-head comparison between different populations is not appropriate due to the differences among the population characteristics (as illustrated in Table 3.1). Additionally, GVAS dataset contains people without diabetes whose glycaemic variability is lower than the diabetic population. For this reason we separately evaluate the performance of the LSTM model for the subjects with and without diabetes as shown in Figure 3.8.

As expected the RMSE values for the diabetic population show higher variability especially in the longer prediction horizons (90 minutes and 120 minutes) comparing with the non-diabetic one, whose variability remains at the same levels for all four prediction horizons. It is worth noting here the effectiveness of the LSTM model to predict future blood glucose values in a non-diabetic population. This is an interesting insight which has not been discussed in many other studies before.

To further quantify the risk of using predicted values of blood glucose we conducted Clark error

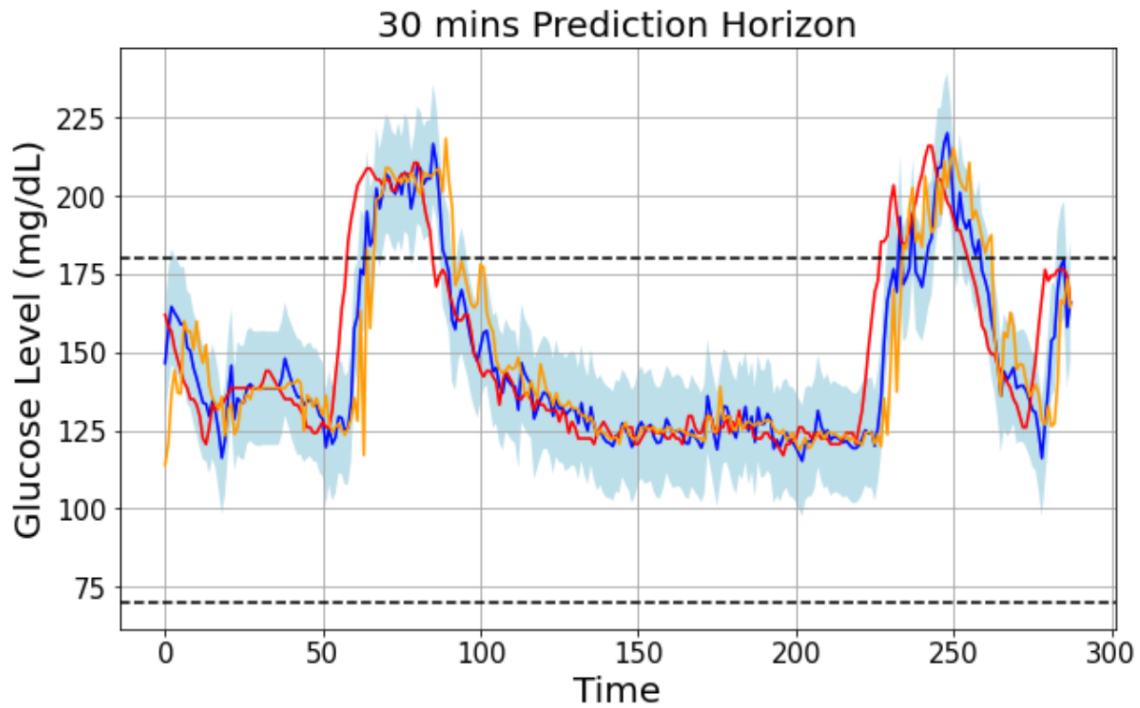


Figure 3.4: 24-hour trajectories of glucose predictions over the 30 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.

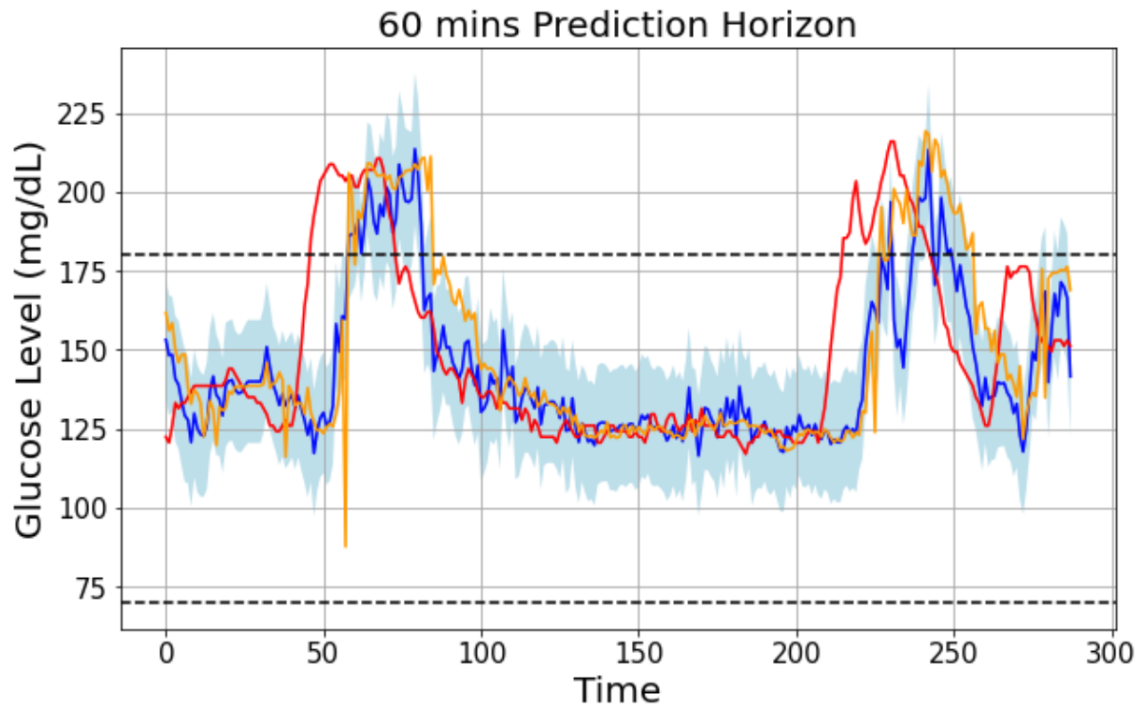


Figure 3.5: 24-hour trajectories of glucose predictions over the 60 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.

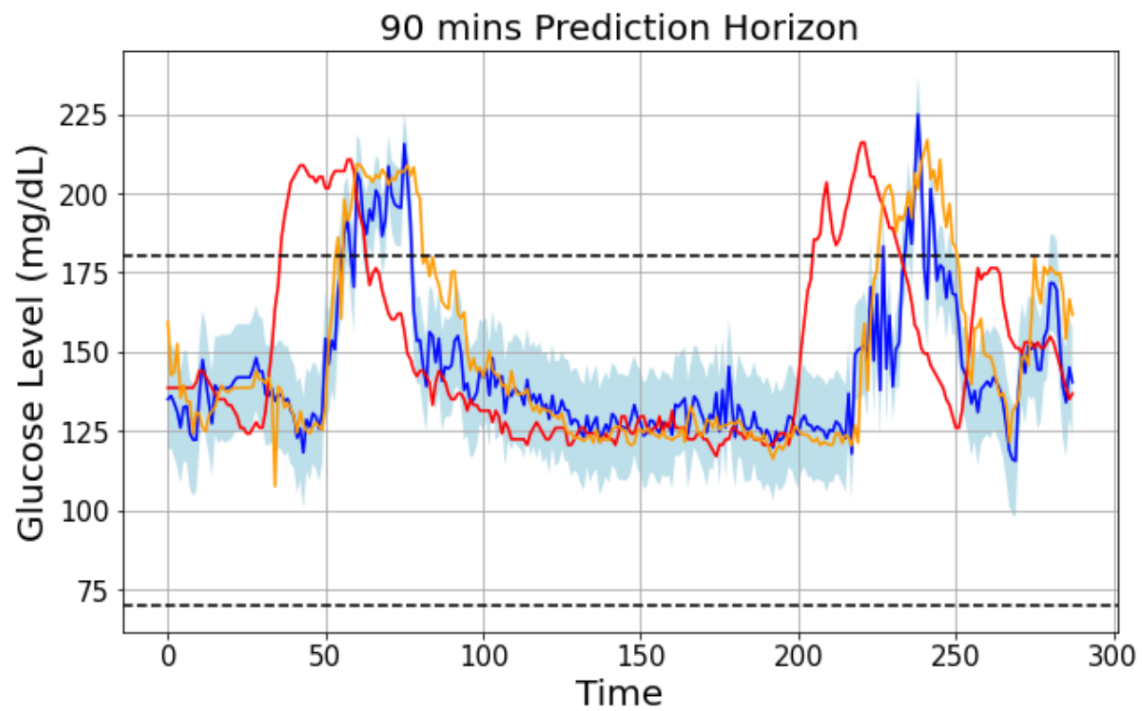


Figure 3.6: 24-hour trajectories of glucose predictions over the 90 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.

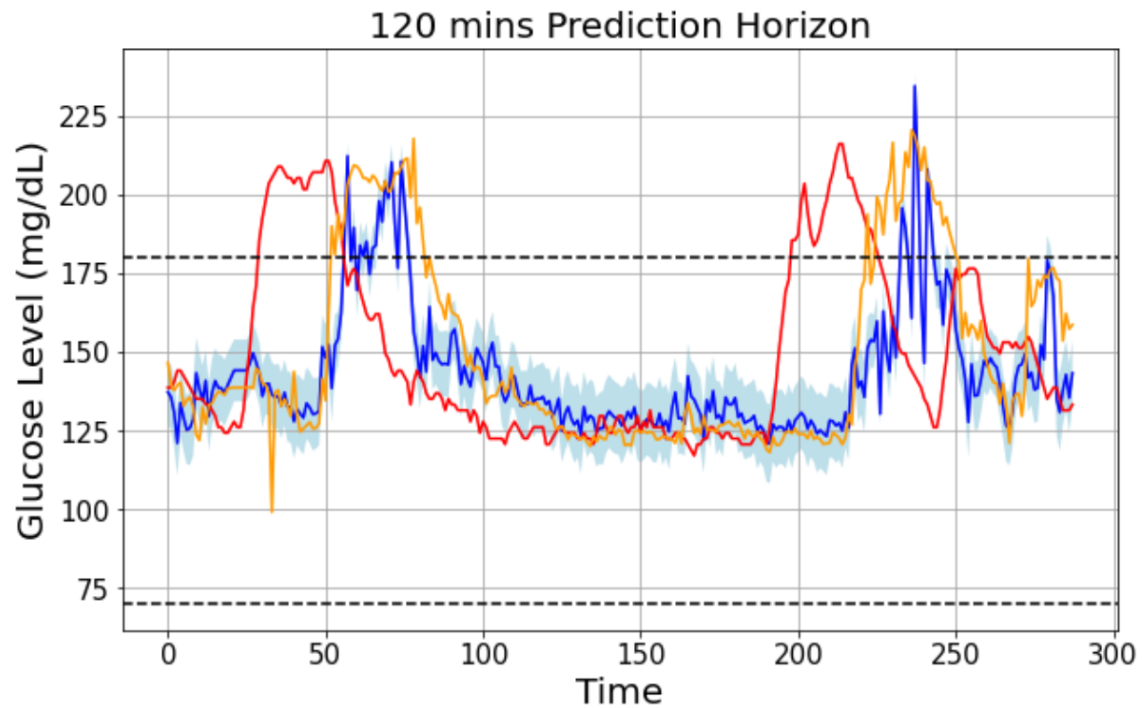


Figure 3.7: 24-hour trajectories of glucose predictions over the 120 minutes prediction horizon. The red line represents actual CGM measurements (288 data points in total), while the yellow and blue lines correspond to the ARIMA and LSTM predictions respectively. The area between the black dotted lines represent the recommended glycaemic range. Above and below are the hyperglycaemic and hypoglycaemic zones.

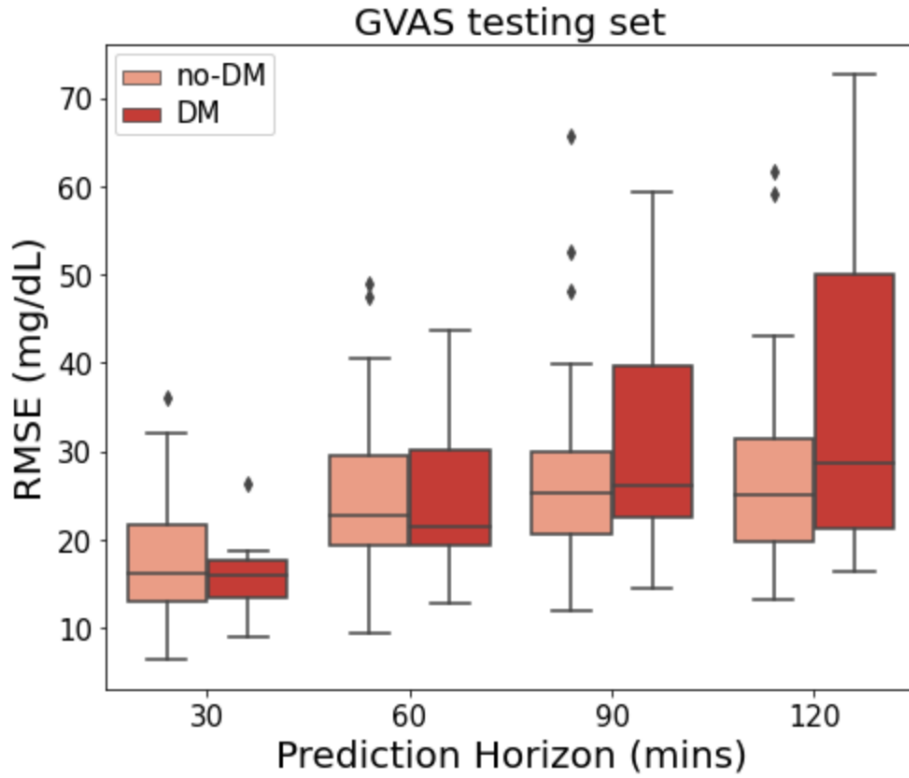


Figure 3.8: Blood glucose prediction performance of the LSTM model on the subjects of the GVAS testing set with diabetes (DM) and without diabetes (no-DM). The lower and upper hinges of each boxplot represent the first (Q1) and third quartiles (Q3), respectively. The median is indicated by the central lines, and the whiskers extend to cover 1.5 times the interquartile range.

grid (CEG) analysis [94, 95, 96] on the GVAS dataset, separating the results between diabetic and non-diabetic patients as shown in Fig. 3.9. CEG analysis, a technique initially for glucose management of diabetic patients, is performed by comparing the CGM values (x-axis) with the estimates (y-axis) derived from the model. As can be observed, the majority of the pair dots for both populations lies between the A and B areas. Specifically, more than 99% of the observations of the diabetic population, and 98% for the non-diabetic population are within CEG zones A and B, signifying that models predictions are accurate or acceptable.

However, an interesting and unexpected observation here is the fact that the non-DM group had twice as much data points in zone D, as opposed to the DM group. This is potentially associated with the hypoglycaemic episodes that people without diabetes seem to have experienced due to stroke. In fact as can be seen in Table 3.1, the percentage of Time Below Range (TBR) for type 2 hypoglycemia (BG levels below 54 mg/dL) is higher in the GVAS dataset, averaging 1.2% compared to 0.4% in the OhioT1DM dataset. This scarcity of hypoglycemic training examples

may limit our model’s ability to make accurate predictions below 54 mg/dL and could also be the reason of the increased percentage of observed-predicted pairs in zone D in Figure 3.9b.

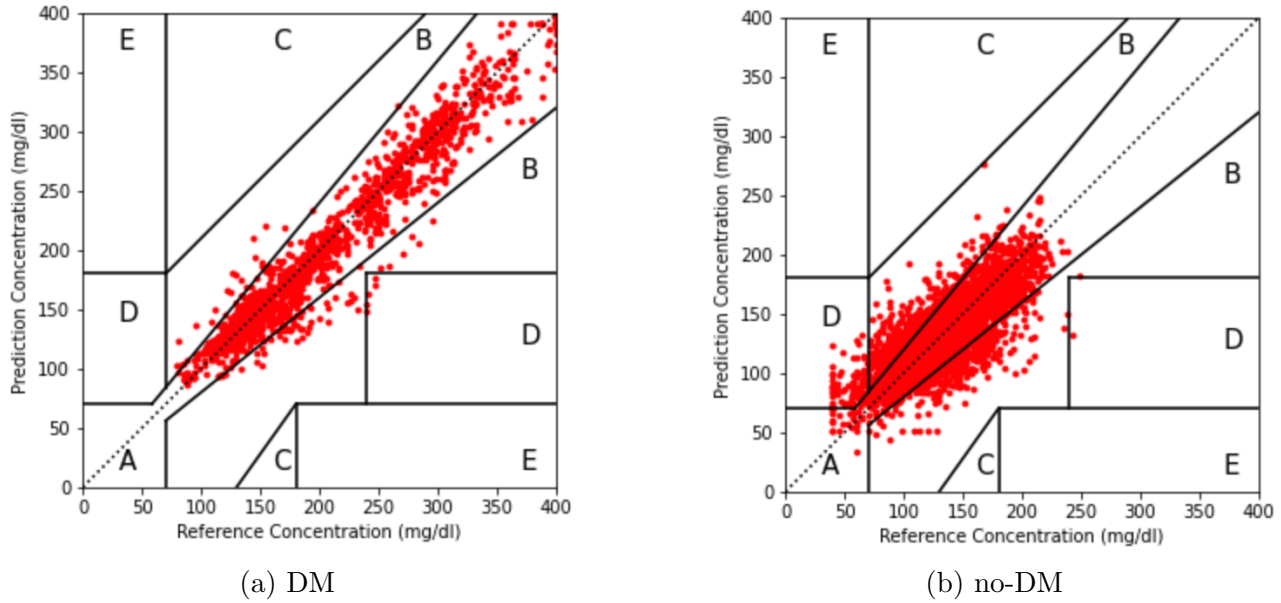


Figure 3.9: Clark error grid (CEG) analysis on the GVAS dataset, showing the comparison between CGM measurements (x-axis) and glucose prediction values (y-axis) for (a) people with diabetes (DM) and (b) people without diabetes (no-DM).

It’s crucial to highlight that, in contrast with numerous existing studies focusing on personalized glucose prediction [40], our approach diverged by excluding meal and insulin (basal or bolus) data from the model input. These data points are prone to noise and increased variability due to inaccurate carbohydrate counting or missing values [38]. Specifically, including these features in the model has led to an increase in the mean RMSE value associated with a larger number of outliers [39].

Choosing to use solely Continuous Glucose Monitoring (CGM) data offers several key advantages. Firstly, it opens up the possibility of embedding the model directly into CGM transmitters, enabling on-device decision support through edge computing [97]. Secondly, it streamlines prediction processes by eliminating the necessity for manual input, thereby avoiding artifacts and facilitating the implementation of automated closed-loop artificial pancreas systems. Lastly, the proposed method proves versatile across a broader spectrum of clinical scenarios, extending its applicability to inpatient populations and the monitoring of conditions related to glucose metabolism, as evidenced by its performance with the GVAS dataset in situations like stroke.

All the above are important and necessary steps towards the deployment of such models in hospital environments, promising an enhancement of patients care in areas such as intensive care units (ICU).

A limitation which will always occur in this kind of studies is the limited amount of available data. Unlike outpatients populations where data can be collected for weeks, even months, inpatients are hospitalised usually for a few days. This is a real burden especially for deep learning models which require a vast amount of data for training and validation. In this work we showed that by borrowing models trained and tuned only in outpatient populations, we can apply them and effectively predict future blood glucose levels for hospitalised patients with or even without diabetes. However, this might not always be the case.

In the literature, considerable fluctuations in glycemic levels are commonly observed, impacting both diabetic and non-diabetic individuals [98]. This becomes particularly evident when subjects employ different insulin delivery methods [99], from multiple daily injections with insulin pens to continuous subcutaneous insulin infusion via insulin pumps. Moreover, other factors like diet, physical activity, and individual metabolism play a significant role in influencing glycaemic dynamics.

In such cases, where the population characteristics are significantly different, there is a possibility that models might not perform equally well. This is the so called zero-shot learning where the model tries to make predictions in example categories that has never seen before. This challenge can be approached as a domain shift problem, considering the new population as a new domain. In fact there are many domain generalization algorithms in various fields such as computer vision [100] where the model is trained to general principles that later help it adapt to new unseen categories more effectively. Hence, such a meta-learning framework for domain generalization can be a very useful approach to tackle this problem in blood glucose level prediction as well [101, 47].

3.5 Conclusion

In this chapter we showed that blood glucose level prediction models, developed in outpatient populations can be effectively used in hospitalised patients with and without diabetes. For the predictions on GVAS dataset we used CEG plots to separately evaluate the model for people with and without the condition. The results highlighted the model's clinical efficacy and ability to provide equally well decision support in both groups of people. We also extended the usual prediction horizon (up to 1 hour in the future) by generating predictions 90 minutes and 120 minutes into the future.

Chapter 4

Prediction of Nocturnal Hypoglycaemia

4.1 Introduction

Nocturnal hypoglycaemia, a phenomenon occurring when blood glucose levels fall below 70 mg/dL, is a very challenging condition in T1DM management. It is particularly challenging for people using a MDI insulin regimen (a combination of slow-acting insulin injections for basal coverage and rapid-acting insulin boluses at mealtimes), since while asleep people are often unaware of the warning signs. Risks factors associated with NH include previous hypoglycaemic episodes, low glycated hemoglobin (HbA1c), years of T1DM, as well as behavioural characteristics such as wrong insulin dosing, insufficient carbohydrate consumption, increased stress levels and late evening physical activity among others. In this chapter, we use data related to some of these risk factors and train ML models in order to predict NH before it happens.

4.2 Data

4.2.1 ABC4D dataset

The main data set used for the development of the algorithms has been collected from a six-month randomized controlled, crossover study [102], aiming to assess the safety and efficacy of an advanced bolus calculator for diabetes (ABC4D) compared with a standard bolus calculator (Clinical Trial Registry No. NCT03963219, ethics approval from the London–Chelsea Research Ethics Committee, reference number: 13/LO/0264). The dataset contains information of 37 adult participants with T1D, which was collected via the ABC4D decision support system. Participants were wearing Dexcom G6 CGM sensors (Dexcom Inc, San Diego, CA, USA), as well as a Fitbit Charge 3 (Fitbit, Inc, San Francisco, CA, USA) activity tracker. In addition, they used the ABC4D smartphone app to record administered insulin (intensified MDI regimen) and meal macronutrient information.

4.2.2 IHART dataset

To enhance the validation of our models, provide evidence regarding their ability to generalize, and evaluate their performance in different glucose monitoring sensor devices, we used the IHART dataset collected from a head-to-head glucose monitoring study (Clinical Trial Registry No. NCT03028220, ethics approval from London–Hampstead Research Ethics Committee, reference number: 15/LO/1679), comparing CGM and flash glucose monitoring [103, 104]. In this study, 40 adult participants, with high-risk of NH, using MDI for their insulin treatment were randomly assigned to CGM Dexcom G5 (Dexcom Inc, San Diego, CA, USA) or FreeStyle Libre 1 (Abbott Diabetes Care, Alameda, CA, USA), intermittently scanned CGM for 8 weeks. An extension phase was conducted after finishing the main study, during which all participants used Dexcom G5 for an additional 8 weeks.

In Table 4.1 we can see an overview of the demographics and clinical characteristics of the populations of the two clinical datasets used in this work. Apart from the age, gender, glucose

monitoring method and insulin regimen of participants, the most common clinical CGM metrics have been calculated as recommended by the international consensus for applications in clinical diabetes care [21]. These include glucose management indicator (GMI) [90], time below range (TBR), time in range (TIR), time above range (TAR), and blood glucose index. The percentage of hypoglycemic nights has also been calculated for both datasets, highlighting the high-risk participants especially in the testing dataset (IHART), with an average of 1 out of 3 nights in hypoglycaemia.

Characteristics	ABC4D	IHART
Age (years)	47 $\pm$ 17	50 $\pm$ 5
Gender (male/female)	10/12	12/8
Mean glucose level (mg/dL)	177.5 $\pm$ 17.6	149.6 $\pm$ 22.6
GMI (%)	7.6 $\pm$ 0.4	7.4 $\pm$ 0.4
eA1c (mg/dL)	138.6 $\pm$ 10.8	133.2 $\pm$ 10.8
TBR (<54 mg/dL) (%)	0.3 $\pm$ 0.3	5.3 $\pm$ 3.7
TBR (54-69 mg/dL) (%)	1.6 $\pm$ 1.2	5.8 $\pm$ 2.8
TIR (70-180 mg/dL) (%)	53.8 $\pm$ 10.8	60.0 $\pm$ 11.0
TAR (181-250 mg/dL) (%)	29.8 $\pm$ 5.7	20.0 $\pm$ 5.8
TAR (>250 mg/dL) (%)	14.4 $\pm$ 7.3	8.7 $\pm$ 7.8
Low blood glucose index	0.6 $\pm$ 0.3	3.0 $\pm$ 1.5
High blood glucose index	10.0 $\pm$ 3.0	6.9 $\pm$ 3.6
Nights with NH (%)	11	34
Glucose monitoring	CGM	flash
Insulin regimen	MDI	MDI

Table 4.1: Demographic and clinical characteristics (*Mean $\pm$ SD*) of the subjects in the two clinical datasets.

4.3 Algorithms

The selection of the algorithms was primarily made by reviewing the best performing models in the literature. As discussed in Chapter 2, the top 3 performances of NH prediction models are tree-based algorithms and SVMs (see Table 2.2). Random Forest’s ability to work efficiently with high-dimensional data, handle missing values and robustness to noise and outliers, along SVM’s robustness to overfitting, and versatility (based on kernel selection) give us the opportunity to experiment with different model configurations and compare then in order to find the best candidate.

4.3.1 Random Forests

Random Forests [105] is a supervised machine learning algorithm which can be used both for regression and classification. It belongs to the category of ensemble learning techniques [106], which essentially combine weaker models together in order to build a more robust model, with lower generalization error. Specifically, whereas decision trees (the fundamental building blocks of Random Forests) on their own are more susceptible to overfitting, their combination in the Random Forests algorithm has shown strong performance in making predictions for various applications.

Random Forests is a modification of bagging (bootstrap aggregation) which builds a large number of decision trees, not correlated to each other, and averages the results. By creating a bootstrapped sample of the original dataset (and equal in size with it), and building decision trees based on a subset of features, Random Forests create a diversified set of identically distributed trees. By averaging the results of each of the individual trees, variance is reduced and the overall model is more robust to noise. The Random Forests pseudocode is presented in Algorithm 1.

Algorithm 1 Random Forests

inputs: T number of trees, N size of the training data, d number of features to be randomly selected out of the total number D (default value of $d = \sqrt{D}$)

for t=1 to T **do**

1. Draw a random bootstrap sample of size N from the training data
2. Grow a decision tree from the bootstrapped data, by repeating the following steps:
 - (i) Randomly select d features without replacement
 - (ii) Split the node based on the feature which maximizes the objective function (i.e. information gain)

end for

output: Use averaging (for regression) or majority vote (for classification) to aggregate the predictions of the trees and assign the predicted value or class label respectively

4.3.2 Support Vector Machines (SVM)

Support Vector Machines [107] are supervised machine learning algorithms which can be used for regression and classifications tasks. SVMs extend on the idea of perceptron [108, 109], but instead of minimizing the misclassification error, they maximize the distance between the decision boundary and the nearest data points from each class. This distance is called margin, the decision boundary is called hyperplane and the nearest points are the support vectors. In other words SVMs try to find the optimal hyperplane which maximizes the margin. A visual representation in two dimensions is shown in Figure 4.1.

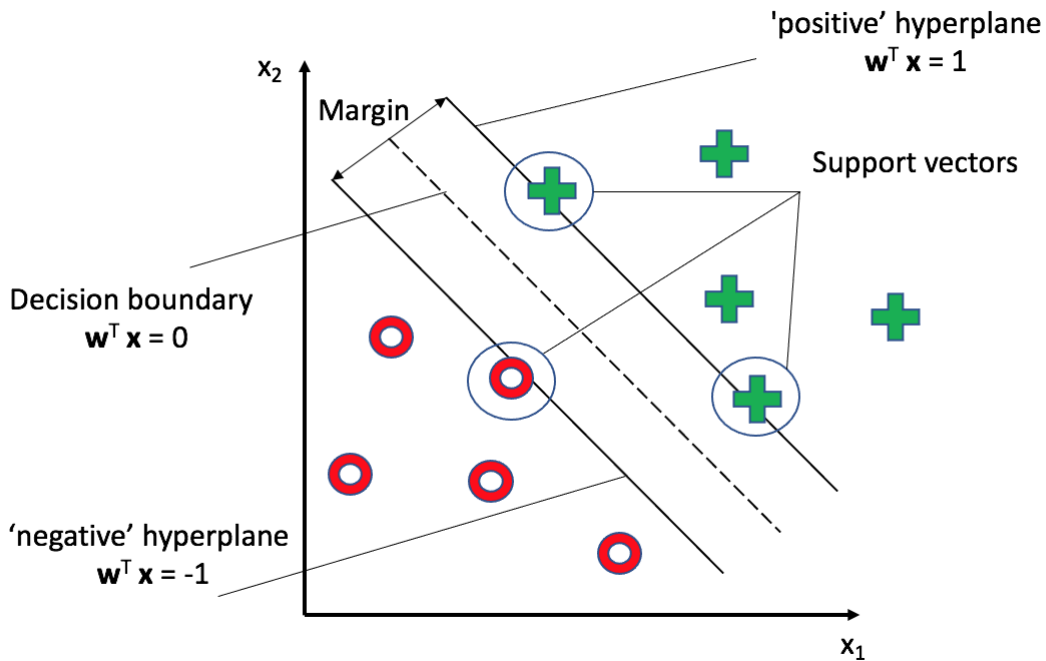


Figure 4.1: SVMs try to find the optimal hyperplane which maximizes the margin

In an ideal world we want to achieve perfect separation between the two classes (in the case of classification). However, in real world data this is almost always not the case. Hence, SVMs try to balance the trade-off between large margins, which often come with more misclassifications, and smaller margins which are prone to overfitting.

Equations 4.1 and 4.2 represent the positive and negative hyperplanes respectively, as shown in Figure 4.1.

$$w_0 + \mathbf{w}^T \mathbf{x}_{pos} = 1 \quad (4.1)$$

$$w_0 + \mathbf{w}^T \mathbf{x}_{neg} = -1 \quad (4.2)$$

By subtracting equations 4.1 and 4.2, we have:

$$\mathbf{w}^T (\mathbf{x}_{pos} - \mathbf{x}_{neg}) = 2 \quad (4.3)$$

If we normalise the equation 4.3 by the norm of $\mathbf{w}$ we get:

$$\frac{\mathbf{w}^T (\mathbf{x}_{pos} - \mathbf{x}_{neg})}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \quad (4.4)$$

The left side of the equation 4.4 is the margin (distance between the positive and negative hyperplane). As mentioned above, the objective of the SVM is to maximize the margin by maximizing the quantity $\frac{2}{\|\mathbf{w}\|}$ subject to the constraints that each point is separated appropriately as follows:

$$w_0 + \mathbf{w}^T \mathbf{x}^{(i)} \geq 1, \quad \text{if } y^{(i)} = 1 \quad (4.5)$$

$$w_0 + \mathbf{w}^T \mathbf{x}^{(i)} < -1, \quad \text{if } y^{(i)} = -1 \quad (4.6)$$

The equations 4.5 and 4.6 restrict all the negative points on the side of the negative hyperplane and all the positive points on the side of the positive hyperplane. However, in reality perfect separation is not achieved and some misclassification occurs. This is called soft-margin classification and is based on the introduction of slack variables ξ which take values from zero to 1, introducing an appropriate cost penalization.

Hence equations 4.5 and 4.6 now become:

$$\mathbf{w}^T \mathbf{x}^{(i)} \geq 1 - \xi^{(i)}, \quad \text{if } y^{(i)} = 1 \quad (4.7)$$

$$\mathbf{w}^T \mathbf{x}^{(i)} \leq -1 + \xi^{(i)}, \quad \text{if } y^{(i)} = -1 \quad (4.8)$$

And consequently the objective to be minimised (subject to constraints) becomes:

$$\frac{1}{2} \|\mathbf{w}\|^2 + C(\sum_i \xi^{(i)}) \quad (4.9)$$

where C is the regularization parameter which controls the trade-off between maximizing the margin and minimizing the classification error. The minimization of the equation 4.9 is performed with quadratic optimization and a pseudocode is presented below:

Algorithm 2 Support Vector Machines - optimization

For a dataset $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where x_i are feature vectors and $y_i \in -1, +1$ are class labels

Objective:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C(\sum_i \xi^{(i)})$$

Subject to:

$$y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)}) \geq 1, \quad \forall i$$

$$\xi_i \geq 0, \quad \forall i$$

4.4 Model Development

4.4.1 Data Preprocessing

Glucose data were gathered and exported as time series, yielding 288 readings per day at five-minute intervals. Any missing values within six consecutive samples (equivalent to 30 minutes) were imputed using linear interpolation. Daily profiles with extended periods of missing data were omitted from the dataset, accounting for less than 5% overall.

Outliers in participants' self-reported insulin values or carbohydrate intake were replaced with the mean or median of each participant's dataset. Additionally, instances where participants consumed a meal after 11 PM were excluded [69], as this was considered an intervention to prevent nocturnal hypoglycemia (rescue carbs).

Flash sensors sample every 15 minutes, unlike CGM sensors, which sample every 5 minutes. To facilitate model evaluation using the test dataset, incomplete daily profiles were discarded as previously described. The flash signal was then resampled to a 5-minute frequency by interpolating intermediate values, aligning it with the CGM time series format (i.e., 288 readings per day).

Labeling of the target class proceeded as follows: For each participant's daily profile, the period from midnight to 6 AM was designated as nighttime. If, during nighttime, glucose levels dropped below 70 mg/dL for at least 20 minutes, it was classified as a hypoglycemic night (Class 1); otherwise, it was labeled as a night without any hypoglycemic episodes (Class 0). Features were extracted from the remaining daytime hours.

4.4.2 Feature Extraction

Glucose data obtained from the Dexcom G6 CGM sensor were utilized to derive glucose-related features for each participant. Within each daytime period, defined from 6 am to midnight, an array of time series characteristics spanning temporal, statistical, and spectral domains were

derived using the TSFEL feature extraction library [110]. Furthermore, conventional metrics related to diabetes management were taken into account, including percentage Time in Range (TIR), percentage Time Below Range (TBR), and Time Above Range (TAR), representing the proportion of time glucose levels remained within, below, or above the target range [70, 180] mg/dL. Additionally, parameters such as high and low blood glucose indexes were considered [111], among others. A comprehensive list of these features is provided in the Appendix.

In order to enhance the capture of CGM signal data, an iterative approach was employed. This involved computing all time series and diabetes-related features for consecutive time windows, beginning from the hour before sleep, extending to two hours before sleep, and continuing up to twelve hours before sleep, as illustrated in Figure 4.2.

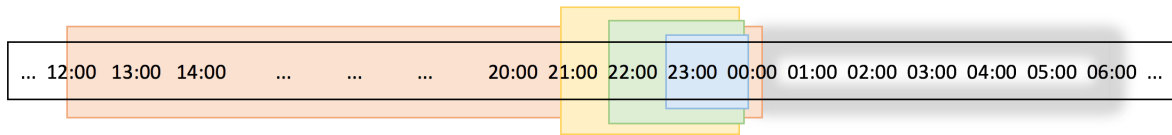


Figure 4.2: Colored boxes represent temporal windows used to extract daytime features (eg, blue: one hour prior to sleep, green: two hours prior to sleep, and orange: 12 hours prior to sleep). The gray shadowed area denotes nighttime.

Data on insulin doses administered by participants was gathered for the creation of the development dataset. To derive insulin-related characteristics, an estimate of the insulin on board model was employed [112]. Likewise, participants manually input information regarding meal intake through the mobile app, including details on insulin boluses. The carbohydrates on board were computed for each participant, indicating the cumulative carbohydrate intake (in grams) from 6 pm until midnight. Additionally, a selection of features was directly extracted from participants' Fitbit trackers, reflecting their daily activity, as detailed in Table 4.2.

4.4.3 Training and Validation

A Sequential Forward Selection method [113] was utilized to pinpoint the most important features. The ensuing set of features is displayed in the table 4.3.

Feature	Metric
Total steps	Total count
Total distance	Meters
Very active distance	Meters
Moderately active distance	Meters
Light active distance	Meters
Sedentary active distance	Meters
Very active minutes	Minutes
Fairly active minutes	Minutes
Lightly active minutes	Minutes
Sedentary active minutes	Minutes
Calories—Total estimated energy expenditure	Kilocalories
Floors	Total count
Calories BMR—Total energy expenditure from BMR	Total count
Marginal calories—Total marginal estimated energy expenditure	Kilocalories
Resting heart rate	bpm

Table 4.2: Description of the Daily Physical Activity Features.

Feature name	Description
Mean_diff	Mean of differences of the signal, calculated for the last hour prior to sleep
Median_diff	Median of differences of the signal, calculated for the last hour prior to sleep
Centroid	Centroid along the time axis, calculated for the last couple of hours prior to sleep
Spectral_distance	Single spectral distance, calculated for the last eight hours prior to sleep
Spectral_decrease	Amount of decreasing of the spectral amplitude, calculated for the last 11 hours prior to sleep
Wavelet_absolute_mean	CWT absolute mean value of wavelet scale, calculated for the last 10 hours prior to sleep
ECDF_Percentile	Percentile values of ECDF, calculated for the last three hours prior to sleep
MFCC	Mel cepstral coefficients, calculated for the last hour prior to sleep
FFT_mean_coefficients	Mean values of each spectrogram frequency, calculated for the last one, two, five, six, eight, nine, 10, and 12 hours prior to sleep
TBR	Percentage TBR, (54, 70) mg/dL, calculated for the last three hours prior to sleep

Table 4.3: List of the Most Relevant Features Derived From the Sequential Forward Selection Algorithm.

The main dataset was divided using a standard 80:20 random split, allocating 80% for training and 20% for holdout. Hyperparameter tuning and effectiveness assessment of the SVM and RF

binary classifiers were performed through grid search within a 10-fold stratified cross-validation (SCV) setup on the training set. An illustration of the data partitioning is presented in Figure A.2 in the Appendix.

The dataset comprises 37 participants and approximately 6000 nights, with 660 nights (11%) experiencing at least one hypoglycemic episode (defined as a minimum of 20 consecutive minutes below 70 mg/dL glucose levels). Due to the imbalanced distribution of the target class, the Synthetic Minority Over-Sampling Technique [114] was employed in each SCV iteration to generate synthetic samples for the minority class. In Figure 4.3 a visual representation of the workflow is presented.

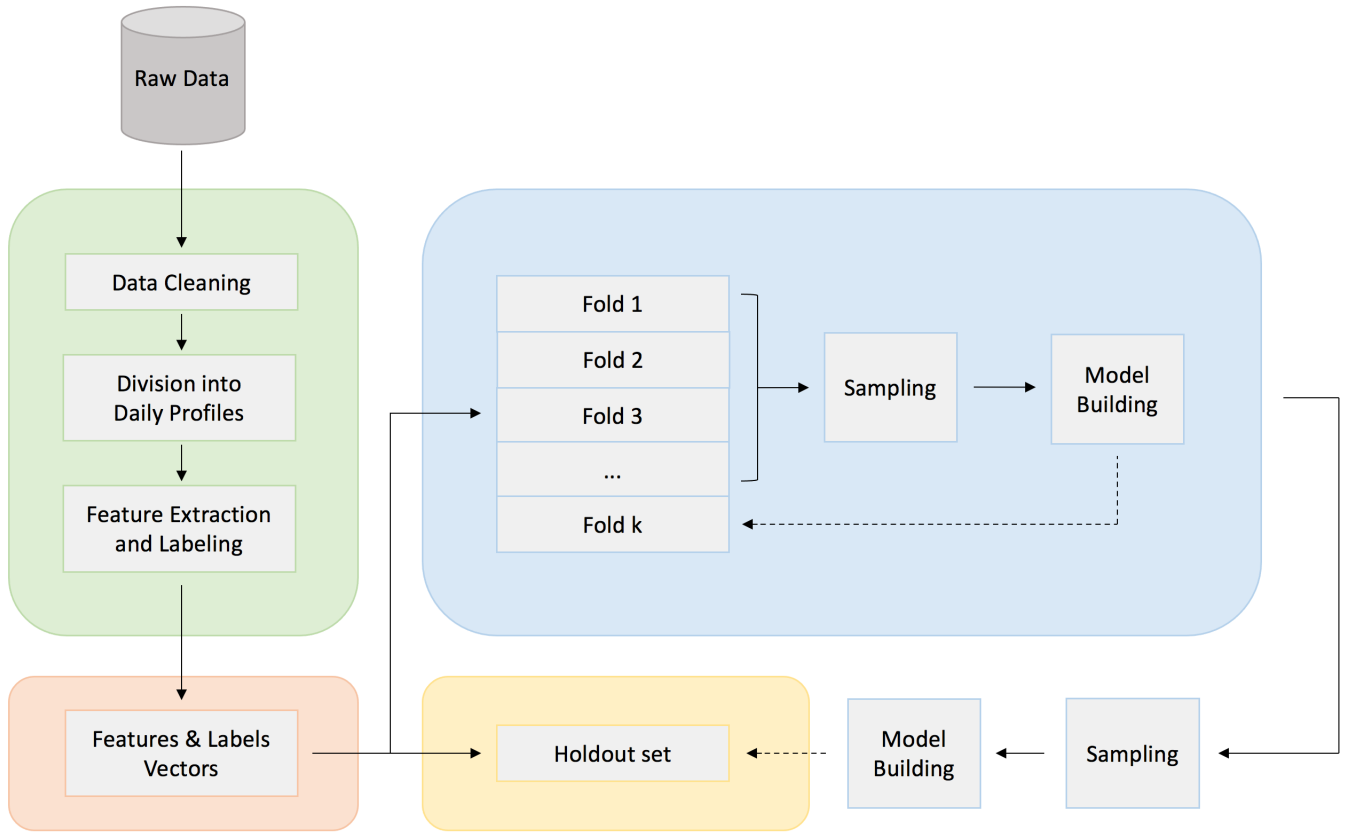


Figure 4.3: Model development and internal validation workflow diagram..

We evaluate the performance of the models through both internal and external validation processes. We compare the two algorithms across both the training and holdout sets of the main dataset. Additionally, we conduct further external validation on the final model using the testing dataset. The primary metric utilized to assess the model's performance is ROC-AUC,

while sensitivity and specificity serve as secondary metrics.

4.5 Results

Table 4.4 displays the performance of both classifiers following a 10-fold cross-validation on the training set and the holdout. Both algorithms demonstrate strong generalization to unseen data. However, SVM notably outperforms the RF classifier, particularly in terms of Sensitivity. Conversely, SVM exhibits balanced performance across all metrics. Additionally, the ROC curve of the SVM algorithm for both the training and holdout sets is illustrated in the Figures 4.4 and 4.5 respectively.

	SVM (kernel: poly, gamma: scale, C:0.1)		Random Forest (n_estimators: 50)	
	Training Set (10-fold CV)	Holdout	Training Set (10-fold CV)	Holdout
ROC-AUC	79.36% (95% CI, 76.86%-81.86%)	78.51%	75.80% (95% CI, 74.00%-77.60%)	76.96%
Sensitivity	73.66% (95% CI, 70.36%-76.96%)	72.13%	42.69% (95% CI, 39.19%-46.19%)	43.88%
Specificity	72.31% (95% CI, 70.51%-74.11%)	71.71%	87.50% (95% CI, 86.60%-88.40%)	89.31%

Table 4.4: The ROC-AUC Score and Sensitivity and Specificity of the Two Machine Learning Classifiers Evaluated During a 10-Fold Cross Validation in the Training and the Holdout Set.

Glucose-related patterns emerged as the most reliable predictors, as illustrated in the Table 4.3, leading to their incorporation into the model. Notably, demographic factors like age, sex, and body mass index did not substantially contribute to NH prediction. This could be attributed to the limited participant pool in the dataset, limiting the range of variation. This outcome was anticipated and doesn't necessarily challenge previous studies [66] that emphasize the significance of demographic features in NH prediction.

The features concerning administered insulin and consumed carbohydrates did not emerge as the most significant either. One possible explanation could be attributed to human error, as participants manually reported insulin units and meal details, resulting in certain discrepancies.

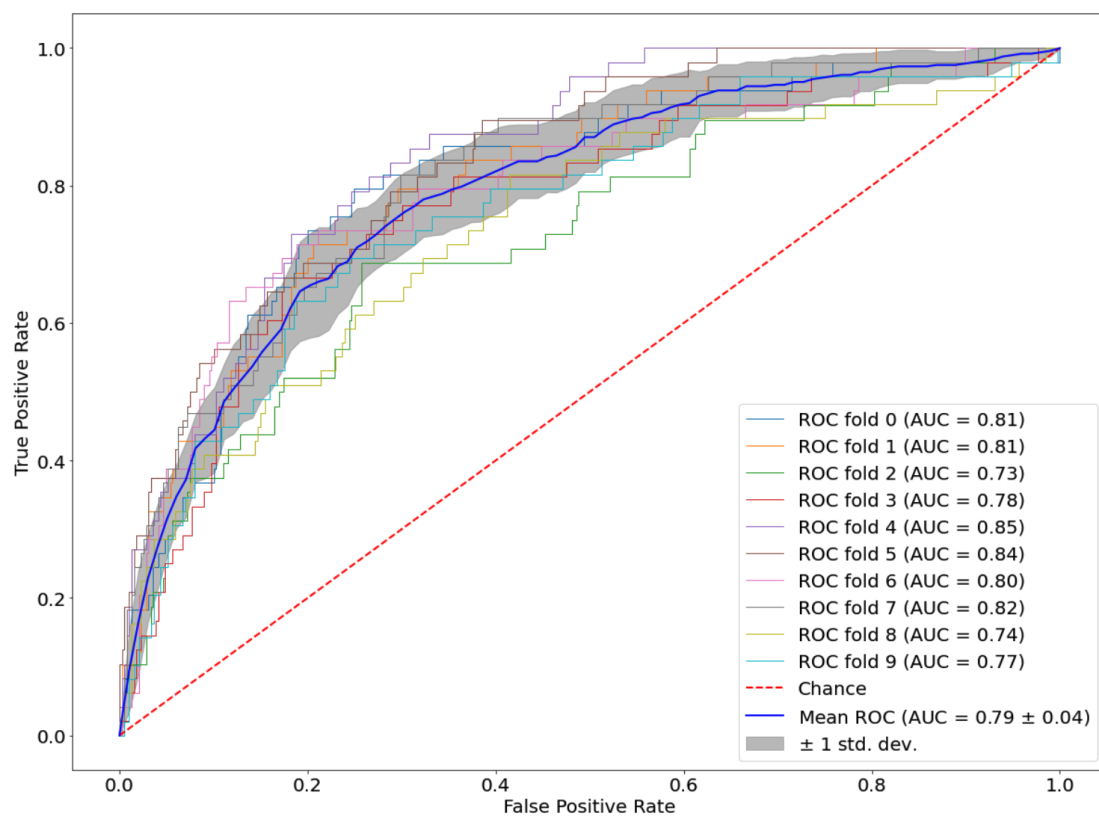


Figure 4.4: ROC curve from the evaluation of the SVM binary classifier, using a 10-fold cross validation in the training set.

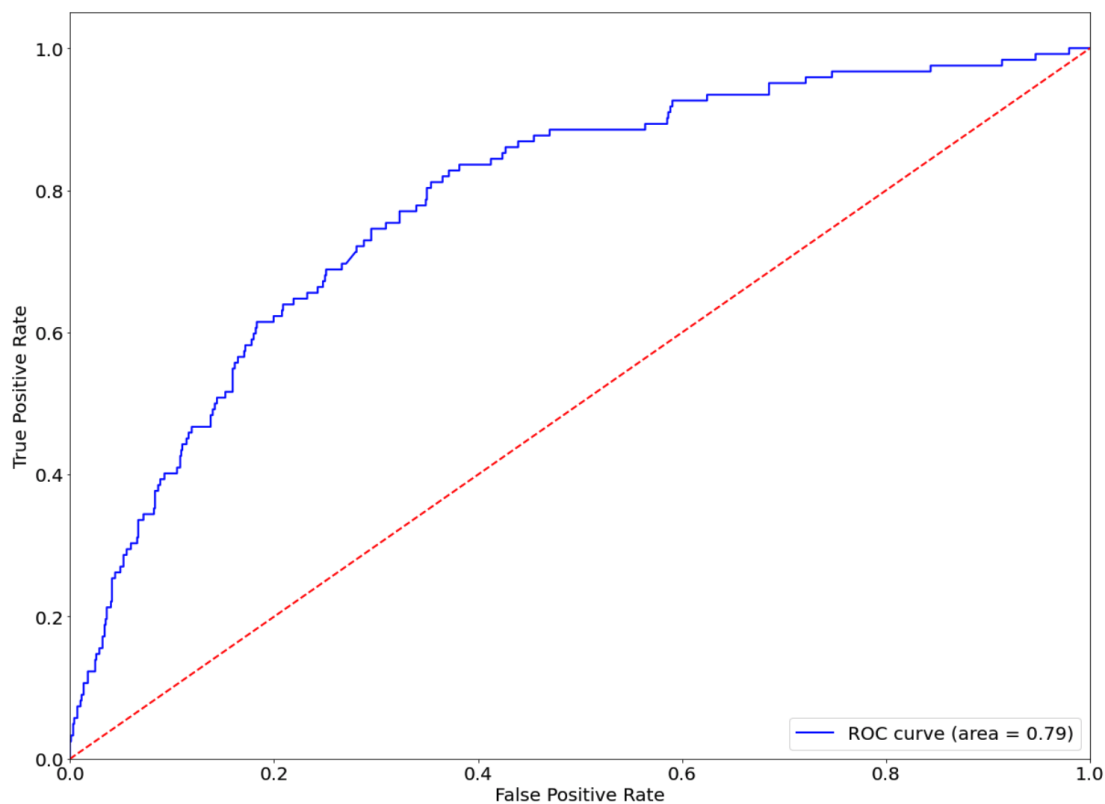


Figure 4.5: ROC curve from the evaluation of the SVM binary classifier on the holdout set.

	CGM (Dexcom G5)	Flash glucose monitoring (Abbott FreeStyle Libre)
ROC-AUC	77.06%	77.74%
Sensitivity	73.76%	82.61%
Specificity	65.63%	56.69%

Table 4.5: Evaluation of the SVM algorithm in a new unseen dataset.

Ultimately, while numerous investigations have underscored the significance of physical activity as a primary risk element in NH events [115] (particularly exercise done in the evening), none of these attributes were chosen by the Sequential Forward Selection algorithm in this study. One plausible explanation could be that the impacts of physical activity, along with insulin and meal data, are already encompassed within the glucose-related features employed by the model.

In Table 4.5, we present the SVM classifier’s performance on a testing dataset comprising a completely unseen population. This group comprises 560 nights (with 189 involving NH), with 20 adult participants wearing a CGM Dexcom G5. The ROC-AUC demonstrates a notably high value, closely aligning with the holdout set, indicating the algorithm’s robustness and capacity to generalize to new data. Additionally, we assessed our model within the same population, this time utilizing FreeStyle Libre for an additional 551 nights (with 187 involving NH). The ROC-AUC performance remains consistently high, underscoring its adaptability to devices from various manufacturers.

In Figures 4.6 and 4.7, the ROC curve and the Recall-Precision curve of the suggested SVM classifier are displayed respectively, assessed using the testing dataset obtained from flash glucose monitoring sensors. It is important to highlight that the classifier was trained to maximize the ROC-AUC score, which means its Precision may not consistently remain high.

4.6 Discussion

This chapter explores a time series feature engineering library [110], designed to capture distinctive signal traits in continuous glucose data. We demonstrate its viability in predicting

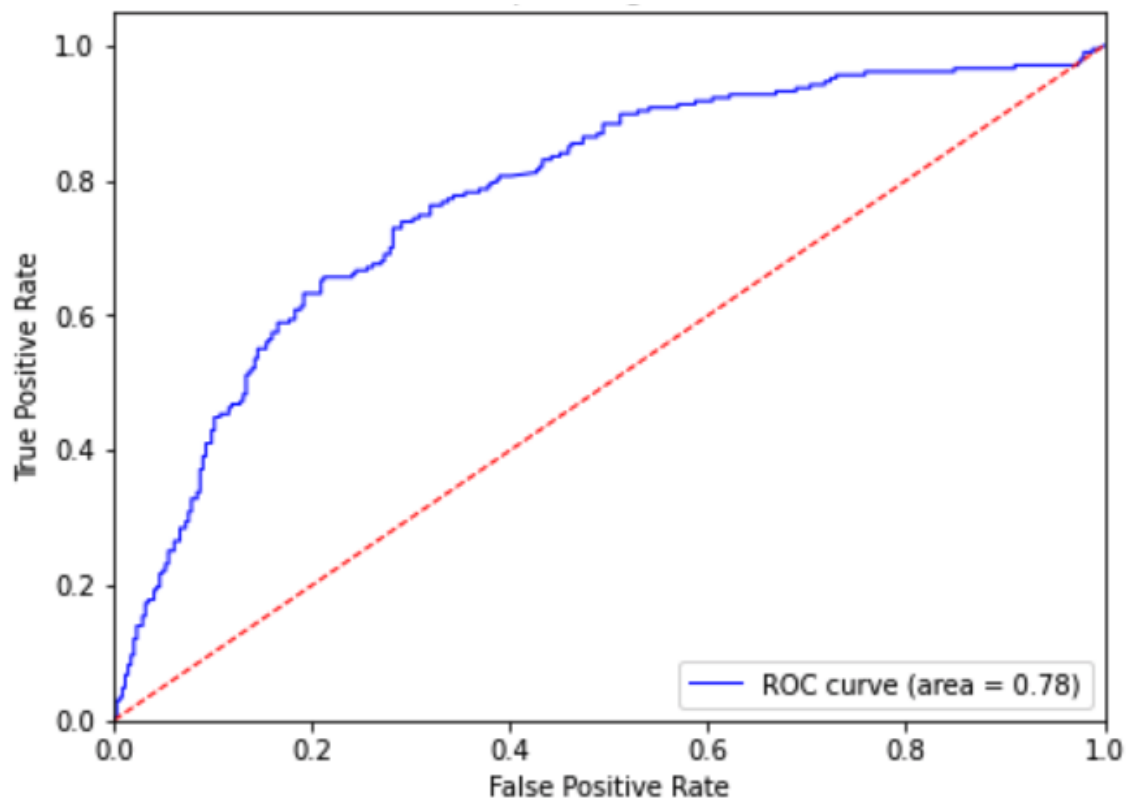


Figure 4.6: ROC curve from the evaluation of the SVM binary classifier on the flash glucose monitoring data.

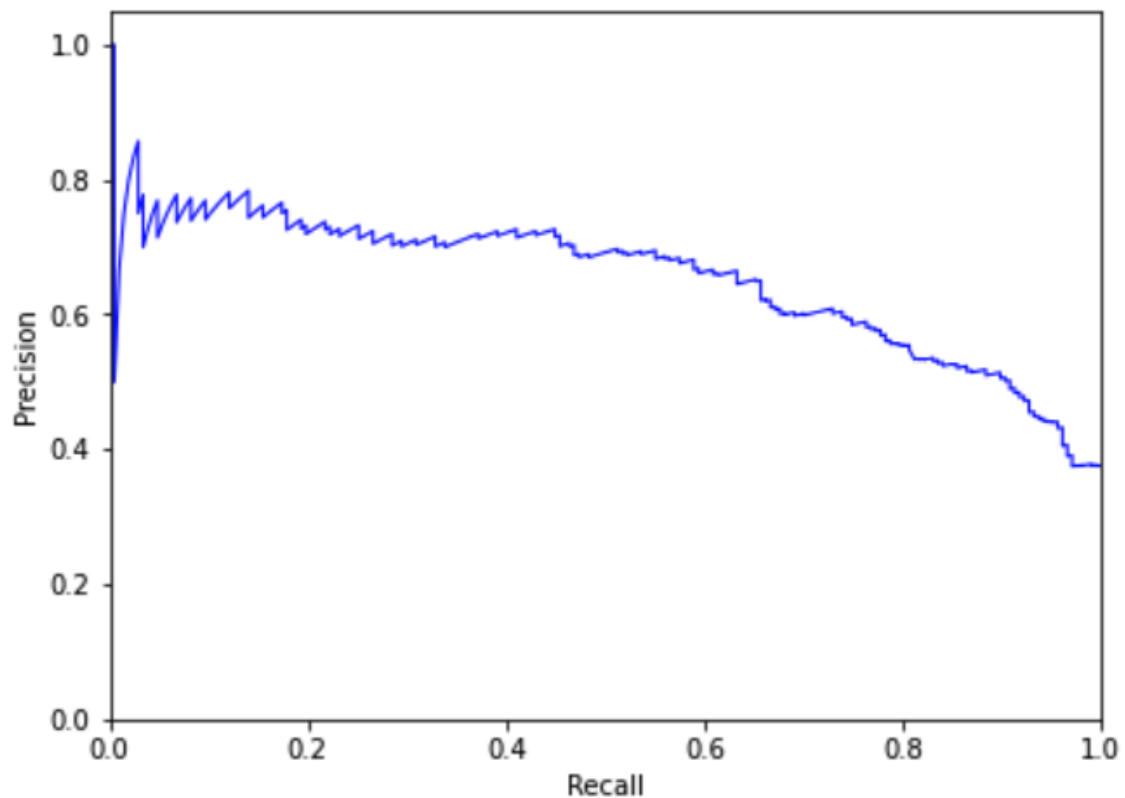


Figure 4.7: Recall-Precision curve from the evaluation of the SVM binary classifier on the flash glucose monitoring data.

nocturnal hypoglycemia (NH) prior to bedtime. Our analysis reveals that the SVM algorithm outperforms RFs in our primary dataset, achieving an AUC score of 79%. Similar findings have been noted in prior research studies as well [68].

We’ve observed that the SVM algorithm relies solely on glucose-related features for its predictions, eliminating the necessity for additional data from sources like administered insulin, meal intake, or physical activity. This stands as a distinct advantage of the proposed model. Specifically, if integrated into an existing CGM/flash system or used as a standalone decision support tool for diabetes management [116, 117, 118], it could offer assistance to individuals with a fear of nocturnal hypoglycemia (NH) by alerting them to potential risks and recommending appropriate actions. This capability is achieved without relying on supplementary data sources (such as meal, insulin, and physical activity records), which may not always be available, complete, or accurate.

The SVM classifier’s generalization capacity underwent further evaluation within a completely new population. Our findings indicate that while its discriminative ability remains relatively high, there is a notable decrease in specificity. One potential explanation could stem from the consumption of rescue carbohydrates prior to bedtime, utilized by participants as an intervention to prevent nocturnal hypoglycemia. Despite excluding such instances during algorithm training, meal data was unavailable in the testing dataset. Consequently, glucose levels may have been influenced by pre-sleep snacks, resulting in a rise in false positives. Additionally, the testing dataset is collected from a population with high-risk of NH, characterized by frequent prior severe hypoglycemic episodes or compromised hypoglycaemia awareness. Specifically, the incidence of nocturnal hypoglycemic events is three times greater compared to the primary dataset.

From the testing dataset we selected the cohort (20 adult participants) which was first assigned to a FreeStyle Libre 1 for eight weeks, followed by a CGM Dexcom G5 for an additional eight weeks. In both instances, the model demonstrated equal accuracy in its discriminatory capabilities, as depicted in Table 4.5. This study represents a pioneering investigation, showcasing that an NH prediction algorithm is adaptable to various sampling frequencies and accuracy char-

acteristics. Furthermore, it illustrates the algorithm’s capacity to perform effectively across sensor devices from different manufacturers, irrespective of the device type. Previous research has underscored uncertainties regarding the role of flash glucose monitoring in managing T1D, particularly for individuals with impaired awareness of hypoglycemia [116, 119, 120]. Specifically, several clinical trials have reported that the use of real time CGM was associated with lower TBR and higher TIR as opposed to intermittent scanned flash glucose monitoring. This has significant clinical implications since both these metrics are among the most important ones for diabetes management as highlighted in the international consensus [21]. Consequently, models like the one proposed in this study offer an avenue for enhancing the capabilities of flash monitoring systems in tandem with CGM counterparts.

Because of the considerable variability in blood glucose response among individuals with T1D, both within and among people, the size of the dataset and the number of observations per individual (such as day-night profiles) are critical factors in algorithm development. A limitation of this study is the relatively small sample size of the development dataset. Despite having access to numerous daily profiles, it’s possible that a single model may not encompass the full range of population characteristics. As a result we may observe that model’s performance is degraded due to changes in the individual glucose patterns (data drift), or due to unseen characteristics of a new population (concept drift).

One solution for the problem above could be retraining, which could aid in understanding individual blood glucose patterns before making overnight predictions. Specifically, the proposed model could be employed during the initialization phase as a way to deal with the cold-start problem (essentially the lack of sufficient and relevant data), by offering a conservative treatment based on clinical guidelines. Following this phase, when significant amount of data have been collected, the model could undergo retraining in order to learn new parameters based on each individual’s new data. That way greater personalization can be achieved by understanding individual blood glucose patterns, and provide more accurate predictions. Finally, a gradual introduction of the retrained model along with close monitoring of its performance will minimize any risk.

4.7 Conclusion

In this chapter, we employ a comprehensive machine learning framework focused on feature engineering to extract features from glucose time series data. Our aim is to build binary classifiers capable of predicting episodes of nocturnal hypoglycemia (NH) prior to bedtime. The model utilizes glucose-related features and achieves an impressive ROC-AUC score of 79%. To validate its effectiveness across diverse populations, we test it on a group of adults with type 1 diabetes (T1D) with high-risk of NH, who were not part of the initial training dataset. While we note a decrease in specificity, the problem at hand is recall-oriented, particularly in identifying NH occurrences broadly. Furthermore, we demonstrate that our model is independent of specific devices, suggesting potential integration with decision support systems to broaden its accessibility, thereby enabling users to be informed about their NH risk proactively.

Chapter 5

Applications and Clinical Trial Design

5.1 Introduction

In this chapter we move from the development and validation of ML algorithms discussed previously, towards real-life applications for diabetes management. Specifically, we use the binary classifier for NH prediction in order to test prevention strategies for NH at bedtime. To do so we conduct an in-silico study based on data from 10 adult diabetes subjects, generated by the T1DM simulator, and highlight the results. Finally, we briefly present the architecture of a decision support system (DSS) necessary to conduct a clinical trial with people in real life, and discuss the design of such a study.

5.2 In-silico study

5.2.1 T1DM Simulator

Type 1 diabetes mellitus simulator (T1DMS), referred to as the UVA/Padova Type 1 Diabetes Metabolic Simulator, is an in-silico diabetes modeling environment based on MATLAB. Initially introduced in 2009 [121], it underwent an update in 2014 [122] and more recently was enhanced

even further [123]. The Food and Drug Administration (FDA) acknowledged its utility in 2018 for the purpose of developing novel strategies for treating type 1 diabetes, serving as an alternative to pre-clinical animal testing. The publicly available version of T1DMS contains a virtual cohort of 30 T1DM subjects (including 10 children, 10 adolescents, and 10 adults with T1DM).

During the simulation, each virtual participant undergoes a daily scenario, offering the user complete control. Participants can be using either open-loop or closed-loop, consume real-world meal scenarios with meal specific timings and quantities, as well as inject insulin boluses (again with specific timings and quantities). At the end of the simulation along with all the generated data, various graphical results are presented to the user. Examples are percentage of time in the different ranges (such as TBR, TIR, TAR) as shown in Figure 5.1, risk indexes (such as low and high blood glucose index) as shown in Figure 5.2, as well as a control variability risk analysis (CVGA) [124], visualising the extreme glucose excursions (2.5th and 97.5th percentiles) of each one of the subjects (Figure 5.3).

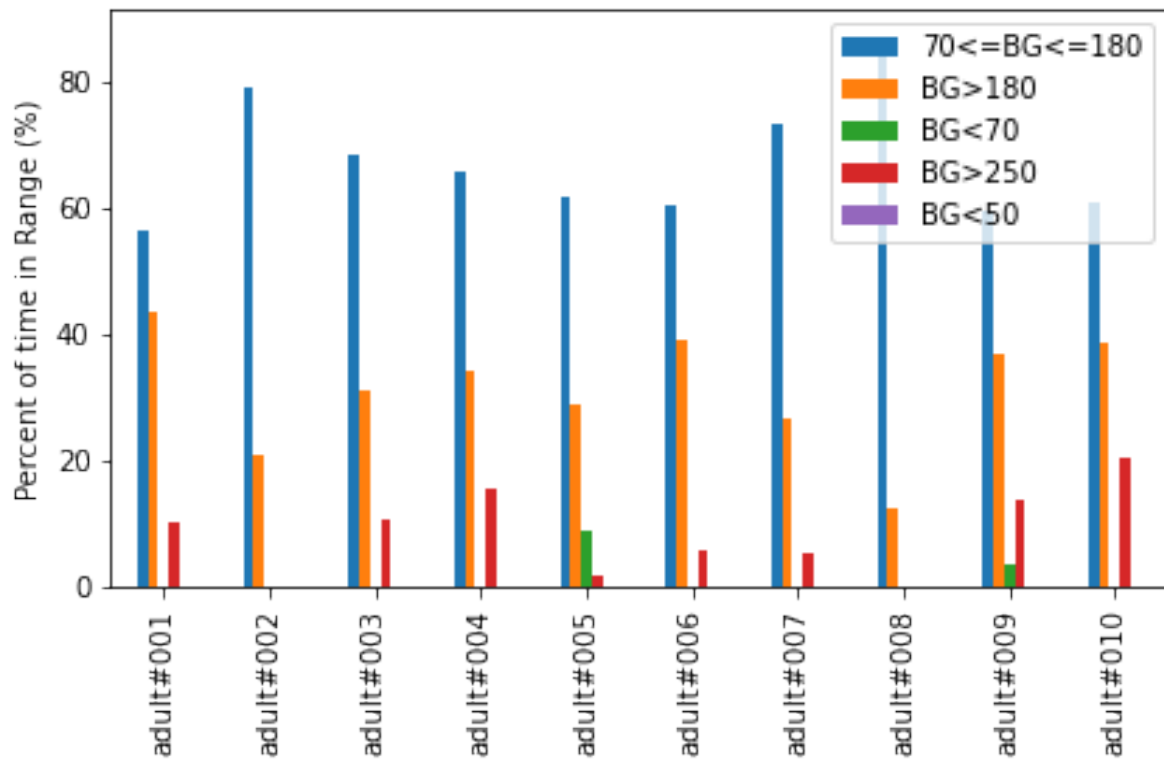


Figure 5.1: Bar chart of the time spent in different blood glucose ranges for each one of the 10 adult participants during a period of 24 hours.

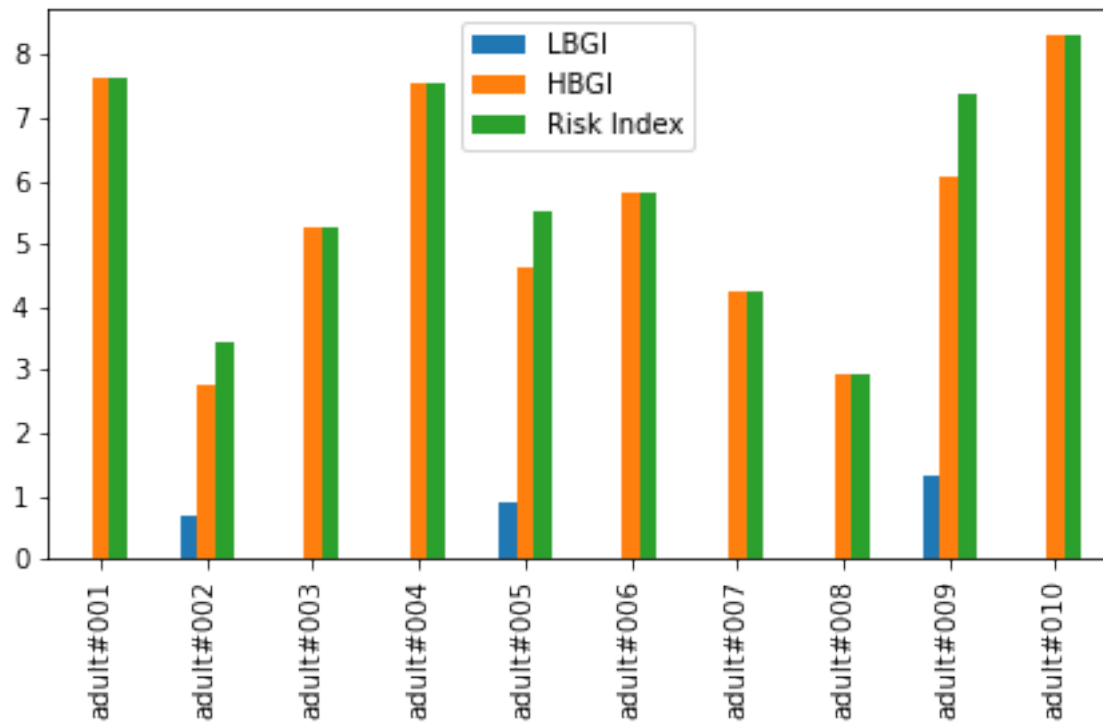


Figure 5.2: Bar chart of the different risk indexes for each one of the 10 adult participants during a period of 24 hours.

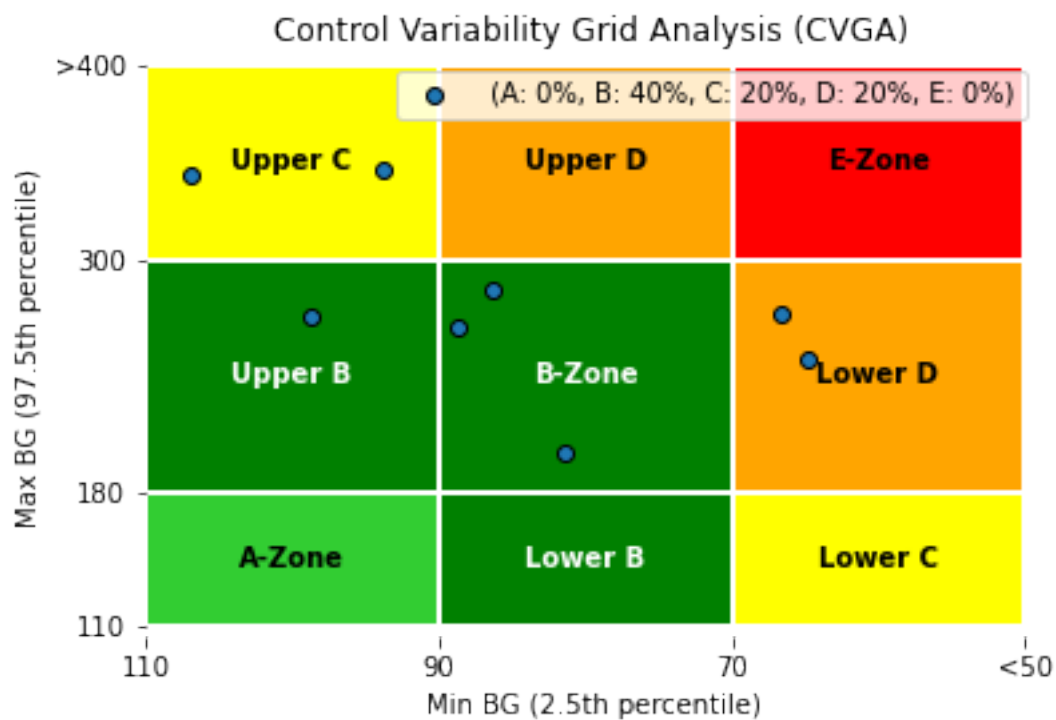


Figure 5.3: CVGA of the 10 adult participants during a period of 24 hours.

CVGA is essentially a tool to evaluate the variability of blood glucose levels across the virtual subjects. It identifies zones which correspond to the quality of euglycaemia with zones A and B being the optimal, and the further up indicating hyperglycaemia-prone and the further right being prone to hypoglycaemia. Zone E in top right corner indicates a very unstable zone with a lot of glucose fluctuations. Figure 5.3 shows the results for 24 hours just as an example, however in reality this analysis is more valuable for a longer period of time.

5.2.2 Study Design

The dataset for the in-silico study contains 365 daily profiles (one year) for each one of the 10 adult participants. The open-loop controller was selected, which combines an external insulin pump with a CGM sensor communicating the glucose readings to the pump using a transmitter. Variability was introduced in the meal protocol of participants and the parameters of the T1DM model [125]. Specifically, we set three meals per day with 15 minutes duration each, and average times for breakfast at 7 am, lunch at 2pm, and dinner at 9pm with a standard deviation of 30 minutes. The meal sizes in terms of carbohydrates was set to be 70 grams, 90 grams, and 130 grams respectively, with a 10% variability. Additionally, carbohydrate miscounting in the range of -30% to +10% was introduced based on a carbohydrate counting study [126], showing that carbs underestimation is more common than overestimation. These configurations remained unchanged during all the different methods and experiments conducted, to guarantee a fair comparison of the results.

Since our focus here is the study of nocturnal hypoglycaemia, we had to make sure sufficient hypoglycaemic nights occur in the data. In a similar research from A. Parcerisas et al [127], the authors mention that they experimented with insulin sensitivity parameters in order to simulate NH. However, in this study we didn't find that the modification of parameters related to insulin sensitivity (such as VMX and kp3) from Dalla Man's Model [122] led to a significant increase in hypoglycaemic nights. On the contrary, a larger meal size for dinner on average, as opposed to breakfast and lunch, proved to be an effective way to introduce more nights with hypoglycaemia in the data since this triggered a higher dose of bolus insulin. However,

an interesting observation is that, on average, the ABC4D dataset from the previous chapter showed a similar amount of carbohydrates consumed by participants during lunch and dinner.

In addition, we restricted any automatic rescue carbohydrate consumption to daytime (6am to midnight) solely, as overnight snacks would act as a treatment for NH. The average percentage time below range (TBR) of participants over 365 nights was around 5%, whereas for the two clinical datasets used in the previous chapter TBR was around 2% for ABC4D and 11% for the IHART study respectively. Additionally, we calculated the TBR during nighttime (between midnight and 6am), which was 7.7% as shown in Table 5.1.

Authors in a previous study [69] have shown that a simple bedtime heuristic for nocturnal hypoglycaemia prediction and prevention is when blood glucose levels fall below 149 mg/dL (8.28 mmol/L). We wanted to test this heuristic against the SVM model developed in the previous chapter as well as against an oracle with absolute knowledge of the future. As such, for every daily profile and for every one of the 10 adult participants, just before bedtime (midnight -6am) three predictions about the upcoming night were made. Specifically, the oracle always predicted with 100% accuracy, the Heuristic approach predicted a hypoglycaemic night if the glucose levels were below 149 mg/dL at midnight and a non hypoglycaemic night otherwise, and the SVM algorithm used the previous 12 hours of CGM data to make a prediction for the night as discussed in the previous chapter.

If the prediction of each one of the tree methods was positive (hypoglycaemic night) we wanted to recommend and evaluate prevention strategies by introducing some rescue carbohydrates (CHO) to be consumed by the virtual subjects. The American Diabetes Association (ADA) recommends the "15-15 Rule" for hypoglycaemia treatment [128]. According to this rule individuals experiencing hypoglycaemia should consume 15 grams of fast acting carbohydrates, wait 15 minutes and check their blood glucose levels again. If they are still below 70 mg/dL they should consume another 15 grams of fast acting carbs, otherwise they should switch to a protein-carbohydrate snack instead. Similarly, the Joint British Diabetes Societies (JBDS) have proposed treating hypoglycaemia with rescue carbohydrates ranging from 10 to 20 grams [129]. They have also proposed a weight-adjusted method in which individuals experiencing

hypoglycaemia should consume 0.3 grams of carbohydrates per kilo of body weight.

However, specifically for this virtual population generated by the T1DM Simulator, Parcerisas et al [127] argue that obtaining 30 grams is the optimal value of rescue carbohydrates for preventing NH. Specifically, they asserted that 30 grams could decrease time below range (TBR) the most (around 35%), while increasing the time in range (TIR) by 1.3%. In this study along with 30 grams of rescue carbs we also test the more conservative approach of 20 grams of carbs as prevention strategies for NH.

5.2.3 Results and Discussion

Table 5.1 shows the in-silico evaluation of various nocturnal hypoglycaemia treatments at bed-time. Based on the predictions from three different methods (an oracle, a simple heuristic and our SVM model), we recommend either a 20 grams or a 30 grams carbohydrate snack at bed-time. It is important to highlight here that the SVM model developed in the previous chapter was not retrained in the simulated data, rather it was used ‘off-the-shelf’. Even though there are differences between training and simulated data, the model’s ability to generalise in unseen populations with different characteristics was previously evaluated. Results show the average TBR, TIR and TAR for the whole day as well as overnight (midnight to 6 a.m.), for the 10 virtual T1DM adults (we chose not to include children and adolescents in order to be consistent with the NH prediction work of the previous chapter which was focused on adult populations). The results of the oracle, heuristic and the SVM algorithm were compared with no intervention. Also, we compare the results of the heuristic method against the SVM algorithm. The statistical significance of the results was determined based on the Wilcoxon signed-rank test.

As can be seen in Table 5.1 participants spent an average of $5.1\% \pm 3.8\%$ of their time below range, with $7.7\% \pm 13.8\%$ of the overnight time in hypoglycaemia. The oracle with both a 20 grams and a 30 grams rescue carbs intervention reduced the TBR to $3.4\% \pm 1.2\%$ and $3.2\% \pm 1.1\%$ respectively. Even though overnight TBR in the case of 30 grams was almost eliminated ($0.2\% \pm 0.7\%$), the more conservative intervention (20 grams) increased TIR (83.8% from 82.1%) while also keeping TAR at the same levels of $12.8\% \pm 8.0\%$.

	No intervention	Oracle		Heuristic		SVM	
		20g	30g	20g	30g	20g	30g
TBR overnight	7.7 ± 13.8	0.8 ± 2.5**	0.2 ± 0.7**	3.4 ± 5.7*	2.8 ± 5.4*	2.5 ± 4.1**,##	2.0 ± 3.6**,##
TBR	5.1 ± 3.8	3.4 ± 1.2**	3.2 ± 1.1**	4.0 ± 2.1*	3.9 ± 2.1*	3.8 ± 1.7**,##	3.7 ± 1.7**,##
TIR overnight	90.3 ± 12.9	97.1 ± 3.5**	97.1 ± 3.0	93.3 ± 5.1	91.4 ± 6.3	94.6 ± 3.7#	93.3 ± 4.1#
TIR	82.1 ± 7.7	83.8 ± 8.4*	83.8 ± 8.4	82.9 ± 7.9	82.3 ± 7.7	83.2 ± 8.0#	82.8 ± 7.9#
TAR overnight	2.0 ± 3.1	2.1 ± 3.1*	2.6 ± 3.0**	3.3 ± 4.0*	5.8 ± 6.7*	2.9 ± 3.6**	4.7 ± 4.7**
TAR	12.8 ± 8.0	12.8 ± 8.0	13.0 ± 7.9**	13.1 ± 7.7*	13.8 ± 7.5*	13.0 ± 7.8*	13.5 ± 7.6**

Table 5.1: In-silico study results of bedtime intervention for nocturnal hypoglycaemia prevention. Statistically significant difference with respect to ‘No intervention’ ($P < 0.05^*$, $P < 0.01^{**}$). Statistically significant difference with respect to ‘Heuristic’ ($P < 0.05^{\#}$, $P < 0.01^{\#\#}$)

The primary goal of NH prevention is the reduction of TBR. However this should not be aimed at any cost. An increase of TAR is a common side effect of the treatment and should also be considered in the evaluation. As opposed to Parcerisas et al [127], where 30 grams proved to be the optimal value of the carbohydrate snack, in this in-silico dataset we found that a more conservative approach (20 grams rescue carbs) seems to have equally good results in NH reduction, while also preventing the unwanted side effects of hyperglycaemia.

The simple heuristic approach, which predicts a hypoglycaemic night if the blood glucose levels fall below 149 mg/dL (8.28 mmol/L) at midnight, reduced the TBR to $4.0\% \pm 2.1\%$ and $3.9\% \pm 2.1\%$ for 20 grams and 30 grams rescue carbs respectively. Similarly, the SVM algorithm also reduced TBR to $3.8\% \pm 1.7\%$ and $3.7\% \pm 1.7\%$. TIR was not impacted significantly in neither methods. On the contrary TAR was increased in all four scenarios. This is often unavoidable as a result of over-treatment due to false positive predictions from the models. In fact it has been observed in other studies as well [69, 127].

Lastly we did a comparison in performance of the heuristic algorithm against the SVM model. As can be seen in Table 5.1, the reduction of TBR and overnight TBR was more efficient with the SVM treatment as opposed to the heuristic, both for 20 and 30 grams of carbs. In fact SVM with 30 grams of rescue carbs achieved the lowest overnight TBR at $2.0\% \pm 3.6\%$ and the lowest TBR at $3.7\% \pm 1.7\%$ among all interventions (excluding oracle). Additionally, even though none of the two methods improved TIR with respect to no intervention, a comparison between them showed that SVM has a positive impact both on TIR and overnight TIR. These results show that the proposed algorithm, if used for NH prevention, outperforms the simple bedtime heuristic by further reducing the TBR without impacting TAR.

In this study we used our SVM algorithm to predict NH in a cohort of 10 virtual adults with T1DM over a period of 365 days. We recommended interventions for NH prevention at bedtime in the form of rescue carbohydrates, and evaluated their effectiveness. The results were quite satisfactory and show that the proposed model could be a viable solution for NH prevention. Other methods could also be explored using the same model. For example in the case of insulin pump users, people can reduce the overnight insulin dose if a hypoglycaemic night is likely to

occur.

One limitation of this study is the small group of subjects in which the interventions were tested. Even though an expanded version of the simulator featuring 300 participants exists, unfortunately it remains inaccessible to the public as of today. Additionally, as discussed in the previous chapter, NH is a challenging issue mainly for people using MDI insulin regimen. The virtual subjects of the T1DM simulator on the contrary are under insulin pump therapy which shifts our target population. Also we know that a bedtime snack containing proteins and fats along with carbohydrates, could also help further in NH prevention especially in the second half of the night as it can prolong its effect in blood glucose regulation [130]. However, the version of the T1DM simulator that we used only supported a carbohydrates model and thus we could not explore other options.

To address some of the above limitations and provide further evidence for the feasibility and efficacy of our method, a meaningful next step would be a small scale in-vivo clinical trial where real people (with the necessary selection criteria) would take part.

5.3 Fundamentals of Clinical Trials

Clinical trials are conducted to assess novel interventions aimed at preventing or treating diseases in humans. These interventions may include medications, medical devices, surgical procedures, behavioral modifications (such as smoking cessation programs), community health initiatives (such as cancer screening programs), or healthcare delivery systems (such as specialized care units for hospital admissions). Clinical trials are considered experiments because the treatments administered to participants are chosen by investigators rather than themselves or their physicians. Randomized clinical trials are typically regarded as the gold standard for determining treatment efficacy, as they are designed to ensure objective and unbiased evaluation of treatment benefits and risks. Regulatory bodies such as the MHRA (Medicines and Healthcare Products and Regulatory Agency) in the UK, the EMA (European Medicines Agency) in EU, and the FDA (Food and Drug Administration) in the US, mandate that drugs and vaccines

must demonstrate efficacy in clinical trials before they can be legally marketed.

5.3.1 Phases of Clinical Trials

The clinical trial phases outline the successive steps required to introduce a treatment into practical use. Originally coming from drug development, these terms have been broadened to encompass various trial types.

Phase 1 trials serve as the initial step in assessing a new intervention in human subjects. Typically involving a small cohort of individuals, these trials may include either healthy volunteers or individuals with the relevant medical condition. The primary objective of phase 1 trials is to establish a tolerable dosage while gathering crucial data on drug metabolism, excretion, and toxicity. These studies often lack a control group.

Phase 2 studies, comprising slightly larger cohorts of up to 100 participants, begin to gather preliminary efficacy data while maintaining a focus on safety and adverse effects. These trials may adopt either controlled or uncontrolled designs.

Phase 3 trials represent the final approval phase for drug evaluation, typically involving 100 or more participants. They aim to comprehensively assess both efficacy and safety, usually employing randomized control designs.

Phase 1 and phase 2 are commonly referred to as the learning phases of clinical trials, while phase 3 is typically known as the demonstration phase. Occasionally, phase 4 studies are conducted post-market approval. These studies evaluate the real-world effectiveness and long-term safety of an intervention, often taking an observational approach, although sometimes incorporating controlled trials.

5.3.2 Types of Clinical Trial Designs

One of the most common type of clinical trial designs is the parallel design. In this setup, individuals are allocated to both groups simultaneously over the same time frame, rather than

solely gathering data on the experimental groups or assigning treatments sequentially. Each participant in a parallel design is assigned to only one treatment group, typically through randomization to eliminate selection bias. Randomization also offers the advantage of creating comparable groups. In a parallel design, statistical comparison involves analyzing a summary of outcome measures between the treatment groups.

An example of a clinical trial with a parallel design is the IHART study (Clinical Trial Registry No: NCT03028220) mentioned in the previous chapter [103]. In this study 40 adults with type 1 diabetes and a high risk of hypoglycaemia, using MDI insulin regimen were recruited. The participants were randomly allocated to CGM ($n=20$) or flash glucose monitoring ($n=20$) for a period of 8 weeks, as shown in Figure 5.4. The study was designed to test superiority between the two devices, with primary outcome the difference in the percentage TBR ($<59\text{mg/dL}$ or $<3.3\text{mmol/L}$) from baseline to endpoint. CGM proved to be more effective (comparing with flask glucose monitoring) in reducing the time in hypoglycaemia in people with impaired awareness of hypoglycaemia.

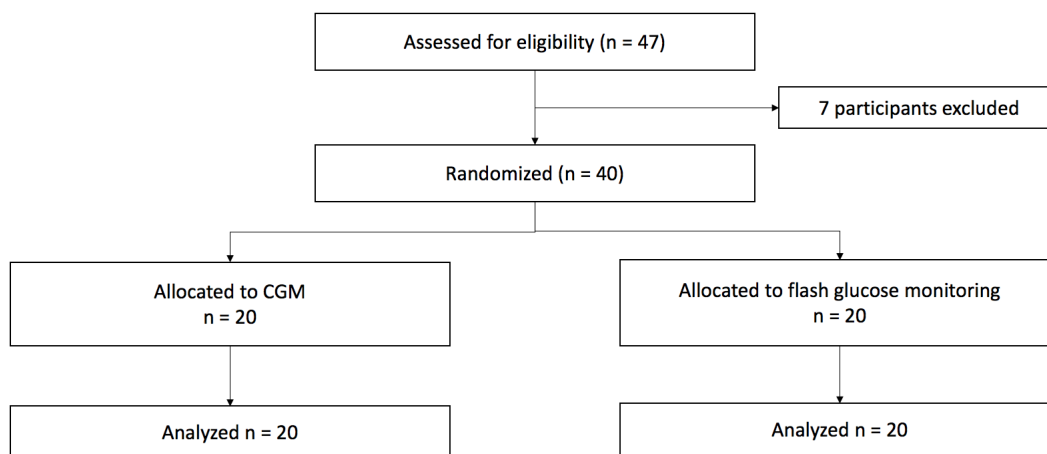


Figure 5.4: Participant recruitment and randomization for the IHART study.

Another type of clinical trial design is the crossover design. In a crossover design the randomization process determines the sequence of treatment administration rather than the allocation of Treatment A or Treatment B to participants. As such, the randomization dictates whether

Treatment A precedes Treatment B or vice versa.

This randomization approach aims to achieve balance between treatment groups and the timing of exposure. A defining aspect of the crossover design is that each participant undergoes both treatments, effectively serving as their own control. This setup is advantageous because variability in outcome measurements tends to be greater between different individuals than within the same individual. By utilizing each participant as their own control, the design accounts for individual-level factors that may influence outcome measurements, thereby reducing variability. As a result, fewer participants are needed to test the hypothesis of interest, making the crossover design more efficient than a comparable parallel design.

Although the efficiency of the crossover design is appealing, it is constrained by several drawbacks. Primarily, it is suitable only for medical conditions and treatments that do not provide permanent cures. Conditions suitable for crossover studies typically involve chronic intensity, where treatments offer symptomatic relief rather than permanent resolution.

Another concern is the potential for carryover effects from one treatment period to another. For example, effects from Treatment A may persist into the period when Treatment B is administered. To address this issue, a washout period is required between treatments to allow any long-standing effects to disappear. The duration of the washout period must accommodate the treatment with the longest potential carryover effect.

An example of a clinical trial with crossover design is the ABC4D study (Clinical Trial Registry No: NCT03963219) which was also discussed in the previous chapter [102]. In this study 37 adults with type 1 diabetes using MDI insulin regimen were recruited. After a 2-week run-in period, participants were randomised to ABC4D (an advanced bolus calculator for diabetes) or control (a standard non-adaptive calculator) for 12 weeks. After a 6-weeks washout period, participants crossed over to control or ABC4D respectively for another 12 weeks equally, as shown in Figure 5.5. The study was designed to test the safety and efficacy of ABC4D, as well as its equivalency/non-inferiority against the standard bolus calculator. The primary outcome the difference in the percentage TIR (70-180mg/dL) change during daytime. ABC4D proved to be safe for providing insulin doses recommendations and had equal effects in % TIR with

the control.

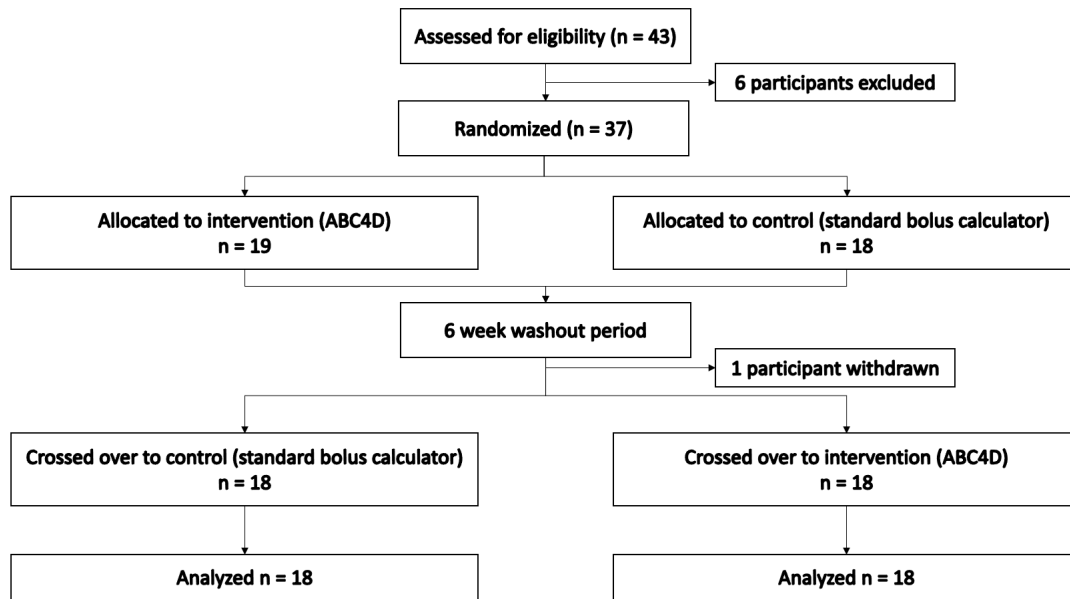


Figure 5.5: Consort flow diagram of the ABC4D study.

There are many extensions and variations of clinical trial designs, based on the specific needs of each study. One example is the group allocation design, where the randomization unit is not a single person but rather a whole clinic or an entire group of individuals. Another variation is the factorial design where two or more experimental interventions are tested simultaneously, often for economical reasons. However in this type of design it is crucial to confirm that there is no interaction between the treatments.

5.3.3 Randomization and Blinding

Randomization refers to the process of assigning participants to different groups (such as treatment and control) in a random manner. It is one of the fundamental principles in clinical trials design and serves several important purposes, including avoiding selection bias and confounding by indication.

Selection bias occurs when systematic differences such as age, the severity of disease, or even clinical centers, exist between the groups being compared. On the other hand, confounding by indication occurs when the reason for prescribing a treatment is associated both with the

treatment received and the outcome being measured. Both biases would distort the comparison between treatments and lead to erroneous conclusions. Randomization helps to mitigate these biases by ensuring that potential confounding variables are evenly distributed across the study groups, and as a result enhancing the validity of the study results.

Blinding or masking, is another critical method in clinical trials design. It is used to ensure that all participants assigned to treatment and control are treated the same throughout the trial, and objectively evaluate their outcomes. The main purpose of blinding is to mitigate information bias.

Information bias occurs when knowledge of the exposure or intervention influences the outcome assessment or reporting. Blinding prevents that from happening by ensuring that the treatment assignment is not known after randomization. Depending on how many people are unaware of the treatment assignment, for example participants, clinicians, or even data analysts, we talk about single, double or triple blinding respectively. By reducing information bias through blinding, researchers can increase the validity and reliability of their findings, leading to more accurate conclusions about the effects of interventions or treatments.

5.4 Towards a Clinical Trial for NH Prevention

The first step to move from the in-silico study towards a real-life clinical trial is the development of a decision support system (DSS). The DSS uses the proposed machine learning model to predict hypoglycaemic nights and recommends interventions for NH prevention. Along with a DSS, a clinical trial protocol highlighting the objectives and outcomes, the design, and participant information among others, is also necessary.

5.4.1 Decision Support System Design

The proposed model for NH prediction from the previous chapter will be incorporated in a smartphone application. This is quite straightforward since the model has been developed

using the Scikit-learn Python package [131], which allows flawless integration with Swift (the native iOS programming language for app development), through the Core ML python package. Additionally, the synchronization of the CGM sensor with the app will be achieved through the use of the relevant application programming interface (API) of the CGM manufacturer, so that blood glucose time series are available to feed the model near real-time (sampled every 5-15 minutes), the same way it has been developed.

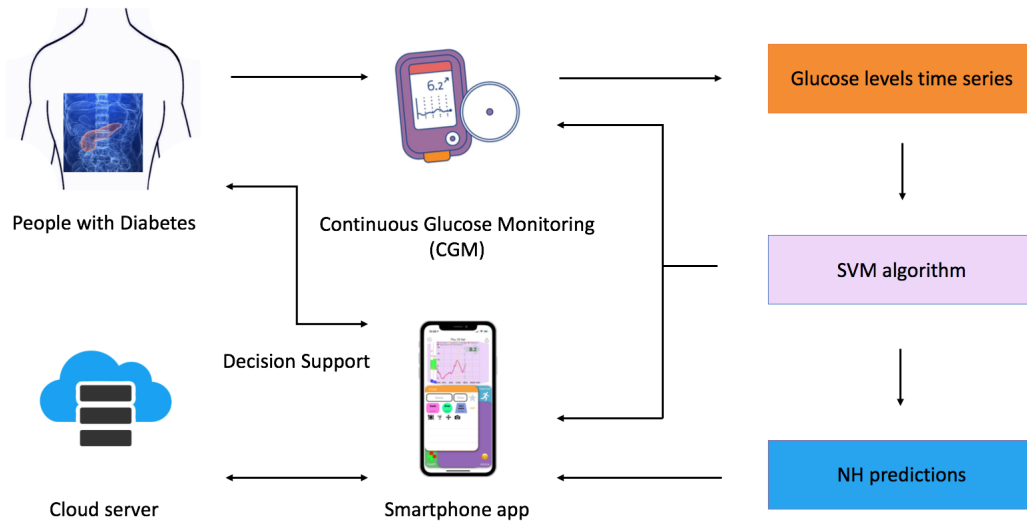


Figure 5.6: Design of a Decision Support System for Nocturnal Hypoglycaemia Prevention.

In terms of the user interface (UI) of the app, the needs for the decision support system are very limited. The user should be able to request a recommendation with the click of a button before going to bed. Based on the predictions made by the algorithm, the recommendation for the upcoming night will be displayed in user's screen. So, for example if the prediction is positive (hypoglycaemic night), the user will receive a message saying it might be a good idea to consume 20-30 grams of carbs before going to bed tonight. Streamed CGM data, model's predictions, recommendations, as well as potential user inputs will be stored in a database in the cloud for future data analysis. A simple diagram of the proposed DSS for NH prevention is presented in Figure 5.6.

The simple design and relatively quick development of the app is a great advantage of the proposed clinical trial. Of course depending on the design of the trial and potential other needs

of the study, more functionalities can easily be added to the app as we discuss in the next section.

5.4.2 Clinical Trial Design and Discussion

Having a DSS in the form of a mobile application, as discussed in the previous section, is mandatory in order to test the effectiveness of our model for NH prevention in real-life scenarios. Specifically, people who are in the treatment group, with a click of a button, will be able to receive recommendations for rescue carbs consumption before bedtime, based on the predictions of the model.

The proposed study design to test the safety and efficacy of a DSS for NH prevention is presented in Figure 5.7. We recommend that the optimal design is a crossover clinical trial design, where each participant will serve as their own control group and thus we will be able to assess more confidently the performance of the treatment. Randomization will be applied to dictate the order with which every participant will be allocated to the two groups (for example, first control and then treatment, or treatment first followed by control). The washout period between the treatments does not have to be too long since there are no long-standing effects in the treatment of NH. In fact the recommendations provided by the DSS regarding carbohydrate consumption, only impact the upcoming night or the next morning at most.

Blinding can also be applied throughout the study, and across everyone involved in the clinical trial (participants, clinicians, data analysts). For that, a duplicate version of the DSS is needed where everything will look exactly the same, but there will be no predictions and recommendations involved at all. So, even if participants ask for a recommendation, the outcome of the DSS will always be that no rescue carbohydrates are needed, as if the prediction for NH was negative. That way no participant or clinician would be able to identify in which group participants belong. However, it is important to consider ethics at this point. It is crucial none of the participants to be at risk, especially if there is some clinical evidence that actions for prevention are available.

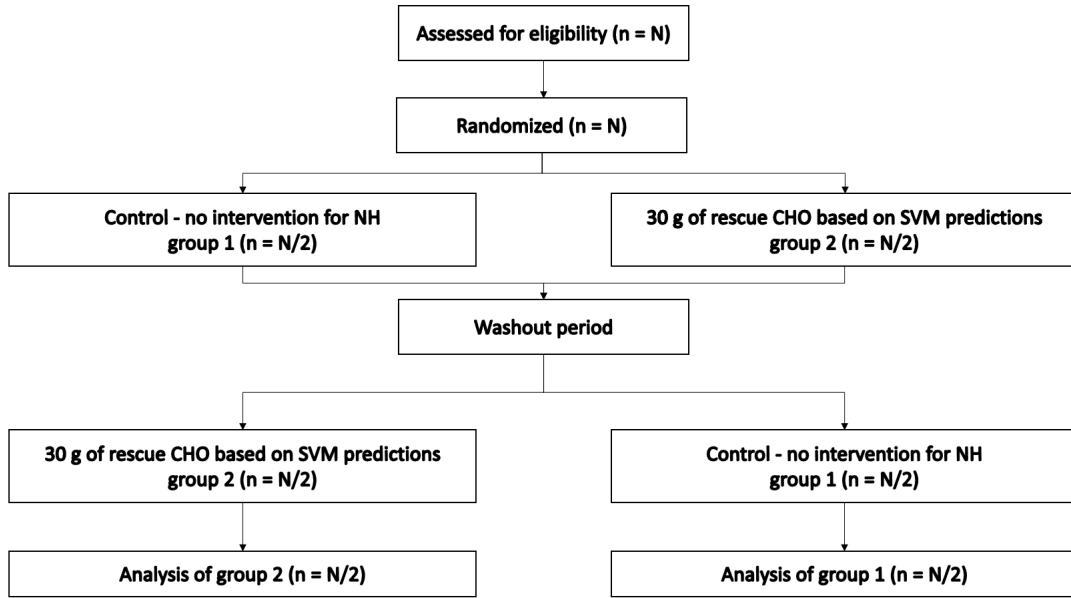


Figure 5.7: Clinical trial design for a NH prevention study.

One solution for the above could be to incorporate a heuristic algorithm (as the one we used in our in-silico study) which will predict NH if blood glucose levels fall below a certain threshold. Essentially, use some clinical guidelines for NH prevention and replicate them in the form of a DSS. That way we can be sure that we take the necessary precautions for all people who are at high risk, while at the same time, increasing the validity and reliability of the outcomes of our study, by keeping blinding across everyone involved. However, making this decision will require more participants, or more time for the study, or both, to observe significant differences in the target metric (here the percentage TBR).

The calculations for the required sample size and study duration are based on the control rates and the expected differences from the interventions. Therefore, if the difference between the control and intervention groups is smaller, a larger sample size and more time are needed to achieve statistically significant results. To illustrate this point we can think of the results from the in-silico study in Section 5.2. The overnight TBR was 7.7% for no intervention, and 2.5% for the SVM algorithm with 20 grams rescue carbohydrates. If we assume the former as the control rate then the effect size (expected difference between control and intervention) was 5.2%. Given these rates, and assuming significance level of $\alpha = 0.05$ and power $(1-\beta) = 80\%$, the required sample size to achieve significance is 219 daily profiles. On the contrary, if instead

of the SVM we had chosen the heuristic algorithm with TBR overnight at 3.4% and an effect size of 4.3%, then the required sample size would have been 348 daily profiles. The formula to calculate the required sample size for binary outcome is shown in equation 5.1. In the formula below, α is the significance level (probability of rejecting the null hypothesis when it is true), $(1 - \beta)$ is the power (probability of rejecting the null hypothesis when is false), and P_1 and P_2 are the main outcome rates for control and intervention respectively.

$$n = \frac{(Z_{\frac{\alpha}{2}} + Z_{\beta})^2(P_1(1 - P_1) + P_2(1 - P_2))}{(P_1 - P_2)^2} \quad (5.1)$$

The eligibility criteria of participants should also be considered. Usually, exclusion criteria comprise of senior people over a certain age, pregnancy, high risk individual due to other chronic conditions, or terminal conditions with a low life expectancy among others. Important inclusion criteria for this study though are considered T1DM for a significant number of years, increasing levels of HbA1c, people experiencing nocturnal hypoglycaemia episodes often, while also people using multiple daily insulin injections. Of course the exact criteria should be decided by the clinical team responsible for the study.

As discussed earlier, one limitation of our in-silico study was the fact that the rescue CHO intervention was only comprised of carbohydrates. In a real-life study, it would be a great opportunity to experiment with a complex of macronutrients, including protein and fats along with the necessary carbs. This can be tested either in the existing intervention group (group 2), or by formulating a third group of people so that the impact of such a snack will also be evaluated against the the carbs only approach.

5.5 Conclusion

In this chapter we presented the results of an in-silico study aiming to prevent nocturnal hypoglycaemia before it happens. We used data from 10 adult virtual subjects with diabetes, simulated using the T1DM simulator. Based on the predictions of an oracle, as simple heuristic

and our own ML model for NH prediction, we recommended a bedtime snack of 20 or 30 grams of carbs as intervention. Results showed that the more conservative approach of 20 grams can reduce TBR quite satisfactory, while also increasing TIR and keeping TAR at the same levels. Both the heuristic algorithm and our proposed model significantly reduced TBR, however our SVM model outperformed the simple heuristic, without impacting TAR. Finally, we present a brief proof of concept for the development of a DSS for NH prevention, and discuss the design of a potential clinical trial.

Chapter 6

Conclusion

6.1 Summary of Results

In the beginning of this thesis we discussed the motivation to do research in the field of artificial intelligence for healthcare, and highlighted the reasons why specifically diabetes is a challenging and important application. In the following a brief introduction on diabetes was provided along with the challenges that people face in order to manage diabetes in their daily lives. Before moving to the technical chapters, a literature review on blood glucose level prediction and nocturnal hypoglycaemia prediction was also presented.

In Chapter 3 we present some evidence that we can predict blood glucose levels of hospitalised patients, with a average RMSE score comparable with that of outpatients. Additionally, a Clark Error Grid analysis showed that the vast majority of glucose level predictions for both people with and without diabetes fall in the areas of either accurate or clinically acceptable results. Finally, LSTM model proved to be effective in predicting glucose levels further in the future (at 90 and 120 minutes prediction horizon) both for people with and without diabetes. As it was expected these predictions had higher variability, especially for the diabetic population.

In Chapter 4 we focused on nocturnal hypoglycaemia prediction. We developed two binary classifiers and showed that SVM outperforms Random Forests, with an AUC score of 79%. It

is important to be noted that the most important features used for the predictions are glucose-related. This is an advantage of the proposed model since theoretically it can be used in an existing CGM/flash system, without the need of external data sources such as insulin, meals and physical activity. The proposed model also showed ability to generalise well, when was externally validated in a new population. Finally, we showed that it is flexible with various sampling frequencies and remains accurate when tested in devices from different manufacturers.

In the last technical chapter (Chapter 5) we conducted an in-silico study aiming to test and compare strategies for NH prevention. The virtual population generated by the T1DM simulator had an average time below range (TBR) of 5% overall and 8% overnight, with no intervention. Rescue carbohydrate snacks were introduced before bedtime based on the predictions of a simple heuristic approach (which predicts NH if glucose levels are below 149 mg/dL) and the SVM model as a way to prevent hypoglycaemia during the night. Both strategies reduced significantly the overnight and the overall TBR. However, while the simple heuristic reduced TBR to 4% overall and around 3% during the night, SVM reduced it even further, achieving 3.7% and 2.2% respectively on average (scores consider both 20 and 30 grams of carbs). Finally, after briefly discussing the fundamentals of clinical trials, we propose a clinical trial for NH prevention in an attempt to test the efficacy and feasibility of the proposed strategies in-vivo.

6.2 Thesis Achievements

The research conducted in this thesis has resulted in several significant accomplishments, including the publication of papers and the presentation of posters at international conferences and in academic journals, as outlined below. Specifically, the work on blood glucose level prediction in Chapter 3 contributed to the first publication in the list in the Journal of Biomedical and Health Informatics (IEEE). Additionally, Chapter 2 is essentially the content of the second publication in the list in the Journal of Diabetes Science and Technology, as well as the first e-poster which was presented in the Advanced Technologies and Treatments for Diabetes (ATTD) conference in Barcelona, Spain (2022). Finally, a side project related to the work on

hypoglycaemia contributed to the last e-poster in the list, which was presented in ATTD in Florence, Italy (2024).

The full list of publications and posters is presented below:

- Zhu T, Afentakis I, Li K, Armiger R, Hill N, Oliver N, Georgiou P. Multi-Horizon Glucose Prediction Across Populations with Deep Domain Generalization. *IEEE Journal of Biomedical and Health Informatics*. 2024 Jul 16.
- Afentakis I, Unsworth R, Herrero P, Oliver N, Reddy M, Georgiou P. Development and validation of binary classifiers to predict nocturnal hypoglycemia in adults with type 1 diabetes. *Journal of Diabetes Science and Technology*. 2023 Jul 11.
- ATTD 2022 e-poster: EP351 / #200 — PREDICTION OF NOCTURNAL HYPOGLYCAEMIA IN ADULTS WITH TYPE 1 DIABETES USING MACHINE LEARNING CLASSIFIERS
- ATTD 2024 e-poster: PD037 / #719 — TIME IN HYPOGLYCAEMIA, BUT NOT COEFFICIENT OF VARIANCE, IS ASSOCIATED WITH INCIDENT RETINOPATHY IN TYPE 1 DIABETES

6.3 Thesis Limitations and Future Work

This concluding section outlines the limitations of this thesis and provides recommendations for future research. The discussion follows the structure of the three main technical chapters, beginning with the challenges and constraints associated with blood glucose level prediction. It then moves on to the prediction and prevention of nocturnal hypoglycemia, highlighting key areas where further investigation is needed. By addressing these limitations, future research can refine and expand upon the findings presented in this work, ultimately contributing to more accurate and reliable diabetes management solutions.

One of the key challenges in developing predictive models for blood glucose levels in hospitalized patients is the limited availability of training and testing data. Unlike outpatients, who may use continuous glucose monitoring (CGM) for extended periods, inpatients typically utilize CGM only during their hospital stay, which is often brief and confined to a controlled setting such as a hospital bed. As a result, there is a scarcity of studies that examine whether machine learning models trained on outpatient data can generalize effectively to inpatient populations. Given the potential differences in physiology, medication use, and activity levels between these groups, gathering more extensive inpatient data is crucial for improving the robustness and reliability of predictive models in this setting.

When a predictive model encounters a population with significantly different characteristics from its training data, its performance may deteriorate due to the domain shift. This issue aligns with the concept of zero-shot learning, where a model attempts to make predictions for cases it has never explicitly encountered before. One potential solution is to frame this challenge as a domain adaptation problem, where the inpatient population represents a distinct domain that requires specialized techniques for better model generalization. In fields such as computer vision, domain generalization algorithms have been developed to help models extract fundamental patterns and adapt more effectively to unseen data distributions. A promising approach for blood glucose prediction could be leveraging meta-learning frameworks for domain generalization, as explored in recent studies [101, 47]. These methods could enable predictive models to better handle variations across different patient populations, thereby enhancing their applicability in inpatient settings.

Another unexplored avenue in blood glucose prediction — both for inpatients and outpatients — is the use of specialized predictive models tailored to different phases of the CGM signal. One of the key challenges in glucose forecasting, particularly during the postprandial phase, is the time lag in CGM readings. To address this, a multi-model approach could be employed. For instance, meal detection algorithms could first identify the onset of the postprandial phase, triggering the use of a specialized predictive model trained on instances of abrupt glucose spikes. This model would be optimized to handle rapid glucose fluctuations more effectively, improving accuracy in this critical period. Once glucose variability stabilizes, the system could transition

back to a general predictive algorithm for standard blood glucose forecasting. By dynamically switching between models based on physiological phases, this approach has the potential to significantly enhance prediction accuracy in both inpatient and outpatient settings.

As discussed in Chapter 4, the variability in blood glucose responses makes both the dataset size and the number of observations per individual critical factors in achieving reliable performance. A key limitation of this study is the relatively small dataset used for model development, which may not comprehensively capture the broad range of individual glucose dynamics across diverse populations. This limitation increases the risk of performance degradation when the model encounters shifts in glucose patterns, such as those induced by new interventions or modifications to existing diabetes management strategies. Additionally, the introduction of previously unseen characteristics in new populations could further impact the model's generalizability.

To mitigate these challenges, one potential approach is to implement a retraining mechanism that allows the model to adapt to an individual's unique glucose patterns before making overnight predictions. This adaptation process would enable the model to account for personal variations in glucose metabolism, leading to more accurate and personalized predictions. Importantly, the retraining procedure can easily be designed to be computationally efficient, requiring only minimal resources. However, several critical questions remain unanswered. For instance, determining the optimal amount of data required for effective model retraining is essential to ensure meaningful learning of new patterns. Additionally, establishing a robust safety release protocol is paramount to protect patients and to guarantee that the retrained model maintains high accuracy and reliability.

Another limitation of the proposed model for nocturnal hypoglycemia prediction is the definition of the night-time period. In this study, we assumed a fixed six-hour window to represent the nocturnal period for all participants. However, this standardized approach may not fully reflect individual differences in sleep schedules and lifestyle patterns, potentially affecting the model's effectiveness. Further research might explore dynamic, personalized night-time definitions to improve prediction accuracy similar to the approaches of a recent study by Herrero et. al. [15]. Additionally, other algorithms such as XGboost (also proposed in the aforementioned

study by Herrero et. al.) and Gradient Boosting Machine (GBM) may offer improvements over the current approach by better capturing intricate patterns in glucose variability.

Another crucial consideration in the development of blood glucose prediction algorithms, including those designed for nocturnal hypoglycemia detection, is the use of more heterogeneous datasets. Incorporating diverse datasets that encompass a wide range of demographic, clinical, and lifestyle variations is essential for improving model generalizability. Furthermore open-source datasets, such as the Ohio T1DM dataset, could be instrumental for providing a standardized benchmark for evaluating and comparing different predictive models. The availability of such datasets could foster collaboration among researchers and facilitate advancements in the field.

As previously noted, nocturnal hypoglycemia (NH) presents a significant challenge for individuals managing Type 1 Diabetes Mellitus (T1DM) through multiple daily injections (MDI) of insulin. However, the virtual subjects utilized in the T1DM simulator are based on insulin pump therapy rather than MDI, making them a slightly different patient population. This distinction is important, as insulin delivery methods can influence glycaemic control and the risk of NH.

Additionally, emerging research suggests that consuming a bedtime snack composed of proteins and fats, in conjunction with carbohydrates, may further mitigate the risk of NH. This dietary approach appears particularly beneficial in stabilizing blood glucose levels during the latter half of the night, reducing the likelihood of hypoglycemic episodes. Unfortunately, the version of simulator used in this study only accounts for carbohydrate metabolism, making it impossible to investigate the potential impact of such dietary strategies.

To address these limitations and enhance the robustness of our findings, the next logical step would be to conduct a small-scale in vivo clinical trial involving real participants who meet predefined selection criteria. Such a trial would allow for direct observation of the intervention's effects in real-world conditions, providing stronger evidence for its feasibility and effectiveness.

Ultimately, the most definitive validation of any predictive model comes from empirical hu-

man data rather than solely from computational simulations. While refining algorithms and conducting benchmark comparisons are valuable components of model development, the true measure of a model's effectiveness lies in its real-world performance. Clinical trials play a crucial role in this process, as they not only validate model predictions but also assess clinical utility, safety, and patient outcomes. A well-designed clinical study such as the one proposed in Chapter 5, would offer critical insights into how the model integrates into standard diabetes management protocols, ensuring that it provides tangible benefits to patients while maintaining safety and efficacy standards.

Appendix A

Supplementary Materials

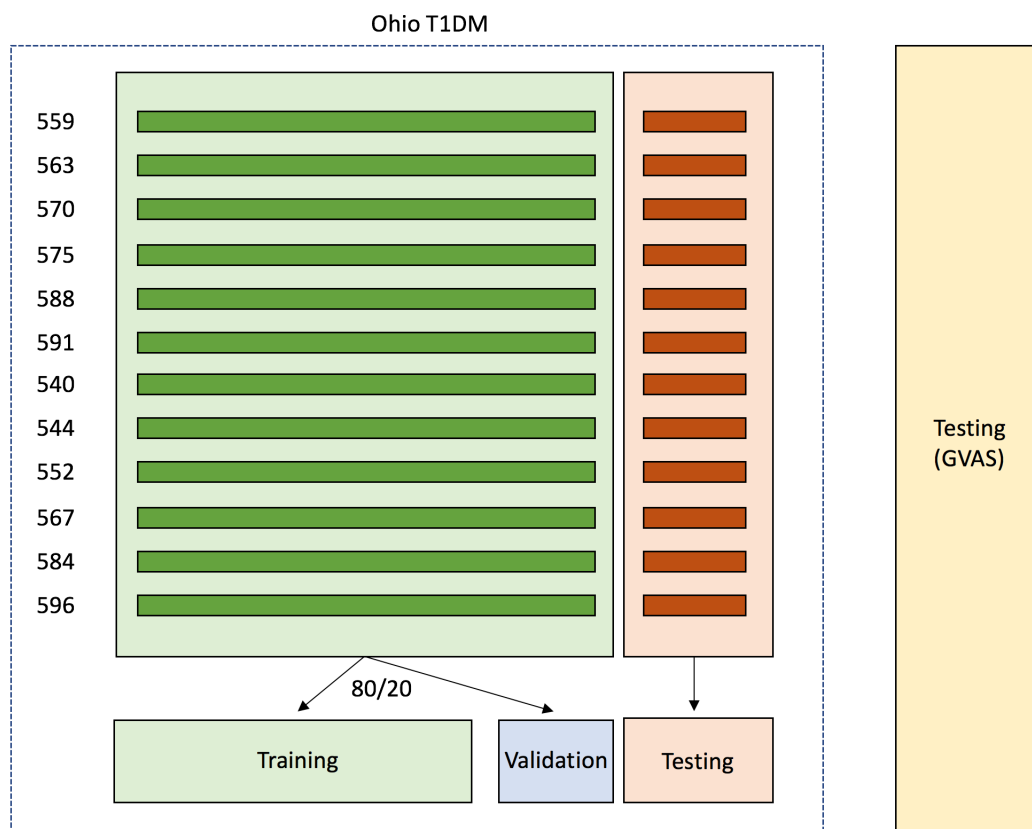


Figure A.1: Visualization of data partitioning for individual and population models. Individual models were trained on the individual training sets and tested on the individual testing sets of the OhioT1DM dataset. For population-level models, we conducted an 80/20 split on the total training dataset, resulting in a training set and a separate validation set. The population models were trained using the actual training sets. Then the validation set was used for feature selection and hyperparameter tuning. Finally, the testing set was used to provide unbiased evaluation and the prediction results reported in this work. Population models were also tested in the GVAS dataset of hospitalized patients without retraining.

Feature name	Description
Time Series Features	
Abs_energy	Absolute energy of the signal
Auc	Area under the curve of the signal (trapezoid rule)
autocorr	Autocorrelation of the signal
Calc_centroid	Centroid along the time axis
Calc_max	Maximum value of the signal
Calc_mean	Mean value of the signal
Calc_median	Median value of the signal
Calc_min	Minimum value of the signal
Calc_std	Standard deviation of the signal
Calc_var	Variance of the signal
distance	Signal traveled distance
ecdf	Values of ECDF along time axis
Ecdf_percentile	Percentile values of ECDF
Ecdf_percentile_count	Cumulative sum of samples that are less than the percentile
entropy	Entropy of the signal using the Shannon Entropy
Fft_mean_coeff	Mean values of each spectrogram frequency
Fundamental_frequency	Fundamental frequency of the signal
hist	Histogram of the signal
Human_range_energy	Human range energy ratio
Interq_range	Interquartile range of the signal
kurtosis	Kurtosis of the signal
lpcc	Linear prediction cepstral coefficients

Table A.1: Full list of glucose-related features for NH prediction model (continued).

Feature name	Description
Time Series Features	
Max_frequency	Maximum frequency of the signal
Max_power_spectrum	Maximum power spectrum density of the signal
Mean_abs_deviation	Mean absolute deviation of the signal
Mean_Abs_diff	Mean absolute difference of the signal
Mean_dif	Mean of differences of the signal
Median_abs_deviation	Median absolute deviation of the signal
Median_abs_diff	Median absolute differences of the signal
Median_diff	Median differences of the signal
Median_frequency	Median frequency of the signal
mfcc	MEL cepstral coefficients
Negative_turning	Number of negative turning points of the signal
Neighbourhood_peaks	Number of peaks from a defined neighborhood of the signal
Pk_pk_distance	Peak to peak distance
Positive_turning	Number of positive turning points of the signal
Power_bandwidth	Power spectrum density bandwidth of the signal
rms	Root mean square of the signal
skewness	Skewness of the signal
slope	Slope of the signal
Spectral_decrease	Amount of decreasing of the spectral amplitude
Spectral_distance	Single spectral distance
Spectral_entropy	Spectral entropy of the signal based on Fourier transform
Spectral_kurtosis	Flatness of a distribution around its mean value
Spectral_positive_turning	Number of positive turning points of the fft magnitude signal
Spectral_roll_off	Spectral roll-off of the signal
Spectral_roll_on	Spectral roll-on of the signal
Spectral_skewness	Asymmetry of a distribution around its mean value
Spectral_slope	Spectral slope
Spectral_spread	Spread of the spectrum around its mean
Spectral_variation	Amount of variation of the spectrum along time
Sum_abs_diff	Sum of absolute differences of the signal
Total_energy	Total energy of the signal
Wavelet_abs_mean	CWT absolute mean value of each wavelet scale
Wavelet_energy	CWT energy of each wavelet scale
Wavelet_entropy	CWT entropy of the signal
Wavelet_std	CWT standard deviation value of each wavelet scale
Wavelet_var	CWT variance value of each wavelet scale
Zero_cross	Zero-crossing rate of the signal

Table A.2: Full list of glucose-related features for NH prediction model (continued).

Feature name	Description
Diabetes related Features	
ri	Risk index
lbgi	Low blood glucose index
hbgi	High blood glucose index
TIR	% Time in range [70, 180] mg/dL
Tbr_1	% time in [54, 70) mg/dL
Tbr_2	% time below 54 mg/dL
Tar_1	% time in (180, 250] mg/dL
Tar_2	% time above 250 mg/dL

Table A.3: Diabetes related features for NH prediction model.

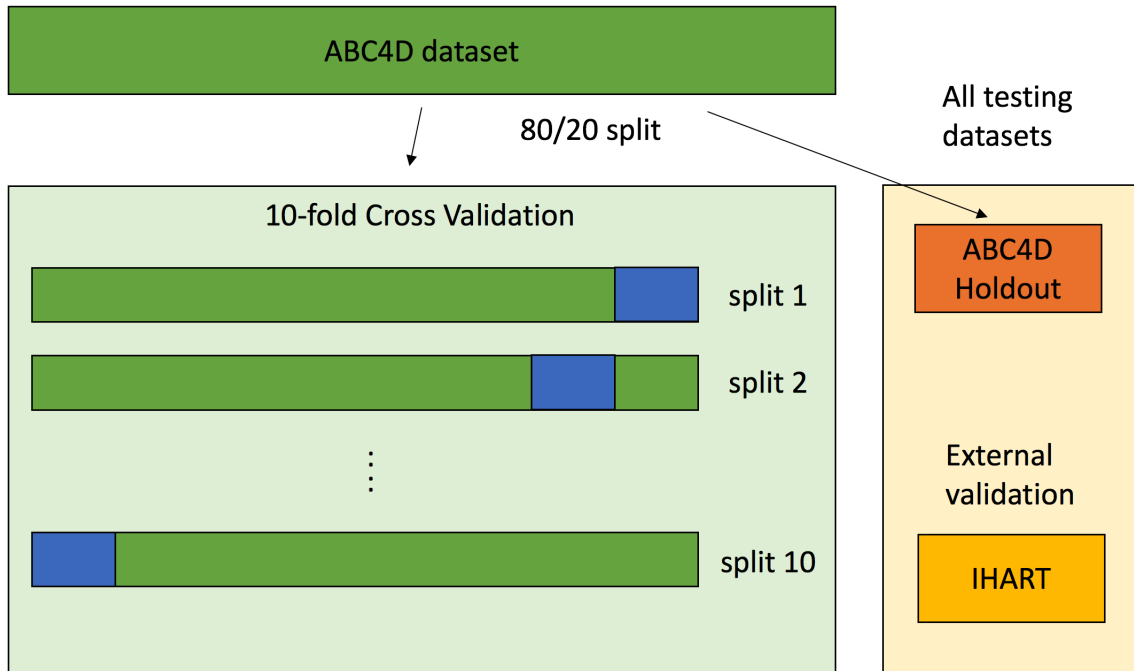


Figure A.2: Visualization of data partitioning strategy utilized in this study for the development of NH prediction models. Initially, we conducted an 80/20 split on the ABC4D dataset, resulting in a training set and a separate hold-out testing set. Following this, a 10-fold cross-validation was performed to facilitate model validation and hyperparameter optimization. To ensure unbiased evaluation, the external IHART dataset was also incorporated.

Bibliography

- [1] J. Bajwa, U. Munir, A. Nori, and B. Williams, “Artificial intelligence in healthcare: transforming the practice of medicine,” *Future healthcare journal*, vol. 8, no. 2, pp. e188–e194, 2021.
- [2] K. Fund *et al.*, “The health care workforce in england: Make or break,” *Kings Fund*.*[Google Scholar]*, 2018.
- [3] P. Dunn, L. Ewbank, and H. Alderwick, “Nine major challenges facing health and care in england,” *The Health Foundation*, 2023.
- [4] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.
- [5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.
- [6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [7] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

- [8] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, “Fast and accurate view classification of echocardiograms using deep learning,” *NPJ digital medicine*, vol. 1, no. 1, p. 6, 2018.
- [9] D. F. Steiner, R. MacDonald, Y. Liu, P. Truszkowski, J. D. Hipp, C. Gammage, F. Thng, L. Peng, and M. C. Stumpe, “Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer,” *The American journal of surgical pathology*, vol. 42, no. 12, pp. 1636–1646, 2018.
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [11] J. Copestake, “Babylon claims its chatbot beats gps at medical exam,” *BBC News*, 2018.
- [12] E. Goh, R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, *et al.*, “Large language model influence on diagnostic reasoning: a randomized clinical trial,” *JAMA Network Open*, vol. 7, no. 10, pp. e2440969–e2440969, 2024.
- [13] B. Levine and A. Brown, “Onduo delivers diabetes clinic and coaching to your smartphone,” *Diatribes* <https://diatribes.org/onduo-delivers-diabetes-clinic-and-coaching-your-smartphone>, 2018.
- [14] B. Kovatchev, A. Castillo, E. Pryor, L. L. Kollar, C. L. Barnett, M. D. DeBoer, S. A. Brown, and N. S. Team, “Neural-net artificial pancreas: a randomized crossover trial of a first-in-class automated insulin delivery algorithm,” *Diabetes Technology & Therapeutics*, vol. 26, no. 6, pp. 375–382, 2024.
- [15] P. Herrero, M. Andorrà, N. Babion, H. Bos, M. Koehler, Y. Klopfenstein, E. Leppäaho, P. Lustenberger, A. Peak, C. Ringemann, *et al.*, “Enhancing the capabilities of continuous glucose monitoring with a predictive app,” 2024.
- [16] W. H. Organization, “The top 10 causes of death,” 2020. Retrieved from: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>.

- [17] I. D. Federation, “Diabetes atlas, 10th edn.,” 2021. Available at: <https://diabetesatlas.org/atlas/tenth-edition/>.
- [18] W. H. Organization, “Global health expenditure database,” 2020. Available at: <https://apps.who.int/nha/database>.
- [19] M. Karamanou, A. Protogerou, G. Tsoucalas, G. Androutsos, and E. Poulakou-Rebelakou, “Milestones in the history of diabetes mellitus: The main contributors,” *World journal of diabetes*, vol. 7, no. 1, p. 1, 2016.
- [20] K. Narayan and F. Fleck, “The mysteries of type 2 diabetes in developing countries,” *Bull World Health Organ*, vol. 94, no. 131, pp. 241–2, 2016.
- [21] T. Battelino, T. Danne, R. M. Bergenstal, S. A. Amiel, R. Beck, T. Biester, E. Bosi, B. A. Buckingham, W. T. Cefalu, K. L. Close, *et al.*, “Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range,” *Diabetes Care*, vol. 42, no. 8, pp. 1593–1603, 2019.
- [22] W. Villena Gonzales, A. T. Mobashsher, and A. Abbosh, “The progress of glucose monitoring—a review of invasive to minimally and non-invasive techniques, devices and sensors,” *Sensors*, vol. 19, no. 4, p. 800, 2019.
- [23] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, S. Garg, L. Heinemann, I. Hirsch, S. A. Amiel, *et al.*, “International consensus on use of continuous glucose monitoring,” *Diabetes care*, vol. 40, no. 12, pp. 1631–1640, 2017.
- [24] J. C. Pickup, “Management of diabetes mellitus: is the pump mightier than the pen?,” *Nature Reviews Endocrinology*, vol. 8, no. 7, pp. 425–433, 2012.
- [25] P. Herrero, P. Georgiou, N. Oliver, D. G. Johnston, and C. Toumazou, “A bio-inspired glucose controller based on pancreatic β -cell physiology,” *Journal of diabetes science and technology*, vol. 6, no. 3, pp. 606–616, 2012.

- [26] N. Oliver, M. Reddy, C. Marriott, T. Walker, and L. Heinemann, “Open source automated insulin delivery: addressing the challenge,” *NPJ digital medicine*, vol. 2, no. 1, p. 124, 2019.
- [27] L. Leelarathna, P. Choudhary, E. G. Wilmot, A. Lumb, T. Street, P. Kar, and S. M. Ng, “Hybrid closed-loop therapy: where are we in 2021?,” *Diabetes, Obesity and Metabolism*, vol. 23, no. 3, pp. 655–660, 2021.
- [28] NICE, “Hybrid closed loop systems for managing blood glucose levels in type 1 diabetes,” 2020. Retrieved from: <https://www.nice.org.uk/guidance/ta943>.
- [29] E. R. Hankosky, D. Schapiro, K. B. Gunn, E. B. Lubelczyk, J. Mitroi, and D. R. Nelson, “Gaps remain for achieving hba1c targets for people with type 1 or type 2 diabetes using insulin: Results from nhanes 2009–2020,” *Diabetes Therapy*, pp. 1–9, 2023.
- [30] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, “Epidemiology of diabetes and diabetes-related complications,” *Physical therapy*, vol. 88, no. 11, pp. 1254–1264, 2008.
- [31] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, “Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series,” *IEEE Transactions on biomedical engineering*, vol. 54, no. 5, pp. 931–937, 2007.
- [32] C. Marling, M. Wiley, R. Bunescu, J. Shubrook, and F. Schwartz, “Emerging applications for intelligent diabetes management,” *AI Magazine*, vol. 33, no. 2, pp. 67–67, 2012.
- [33] R. Bunescu, N. Struble, C. Marling, J. Shubrook, and F. Schwartz, “Blood glucose level prediction using physiological models and support vector regression,” in *2013 12th International Conference on Machine Learning and Applications*, vol. 1, pp. 135–140, IEEE, 2013.
- [34] F. L. Schwartz, J. H. Shubrook, and C. R. Marling, “Use of case-based reasoning to enhance intensive management of patients on insulin pump therapy,” *Journal of Diabetes Science and Technology*, vol. 2, no. 4, pp. 603–611, 2008.

- [35] K. Plis, R. Bunesco, C. Marling, J. Shubrook, and F. Schwartz, “A machine learning approach to predicting blood glucose levels for diabetes management,” in *Workshops at the Twenty-Eighth AAAI conference on artificial intelligence*, Citeseer, 2014.
- [36] E. I. Georga, V. C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni, D. Polyzos, and D. I. Fotiadis, “Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression,” *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 71–81, 2012.
- [37] E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, “A predictive model of subcutaneous glucose concentration in type 1 diabetes based on random forests,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2889–2892, IEEE, 2012.
- [38] S. N. Mehta, N. Quinn, L. K. Volkening, and L. M. Laffel, “Impact of carbohydrate counting on glycemic control in children with type 1 diabetes,” *Diabetes Care*, vol. 32, no. 6, pp. 1014–1016, 2009.
- [39] C. Mosquera-Lopez and P. G. Jacobs, “Incorporating glucose variability into glucose forecasting accuracy assessment using the new glucose variability impact index and the prediction consistency index: An lstm case example,” *Journal of Diabetes Science and Technology*, vol. 16, no. 1, pp. 7–18, 2022.
- [40] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Deep learning for diabetes: A systematic review,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2021.
- [41] S. G. Mougiakakou, A. Prountzou, D. Iliopoulou, K. S. Nikita, A. Vazeou, and C. S. Bartsocas, “Neural network based glucose-insulin metabolism models for children with type 1 diabetes,” in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3545–3548, IEEE, 2006.

- [42] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, “GluNet: A deep learning framework for accurate glucose forecasting,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 414–423, 2020.
- [43] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz, “Using lstms to learn physiological models of blood glucose behavior,” in *2017 39th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pp. 2887–2891, IEEE, 2017.
- [44] A. Aliberti, I. Pupillo, S. Terna, E. Macii, S. Di Cataldo, E. Patti, and A. Acquaviva, “A multi-patient data-driven approach to blood glucose prediction,” *IEEE Access*, vol. 7, pp. 69311–69325, 2019.
- [45] Q. Sun, M. V. Jankovic, L. Bally, and S. G. Mougiakakou, “Predicting blood glucose with an lstm and bi-lstm based deep neural network,” in *2018 14th Symposium on Neural Networks and Applications (NEUREL)*, pp. 1–5, IEEE, 2018.
- [46] S. H. A. Faruqui, Y. Du, R. Meka, A. Alaeddini, C. Li, S. Shirinkam, and J. Wang, “Development of a deep learning model for dynamic forecasting of blood glucose level for type 2 diabetes mellitus: secondary analysis of a randomized controlled trial,” *JMIR mHealth and uHealth*, vol. 7, no. 11, p. e14452, 2019.
- [47] T. Zhu, I. Afentakis, K. Li, R. Armiger, N. Hill, N. Oliver, and P. Georgiou, “Multi-horizon glucose prediction across populations with deep domain generalization,” *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [48] J. Chen, K. Li, P. Herrero, T. Zhu, and P. Georgiou, “Dilated recurrent neural network for short-time prediction of glucose concentration.,” in *The 3rd International Workshop on Knowledge Discovery in Healthcare Data in the 27th IJCAI-ECAI*, pp. 69–73, 2018.
- [49] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, “Blood glucose prediction with variance estimation using recurrent neural networks,” *Journal of Healthcare Informatics Research*, vol. 4, pp. 1–18, 2020.

- [50] T. Yang, R. Wu, R. Tao, S. Wen, N. Ma, Y. Zhao, X. Yu, and H. Li, “Multi-scale long short-term memory network with multi-lag structure for blood glucose prediction,” *KDH@ ECAI*, vol. 45, pp. 136–140, 2020.
- [51] C. Midroni, P. J. Leimbigler, G. Baruah, M. Kolla, A. J. Whitehead, and Y. Fossat, “Predicting glycemia in type 1 diabetes patients: experiments with xgboost,” *heart*, vol. 60, no. 90, p. 120, 2018.
- [52] R. Bevan and F. Coenen, “Experiments in non-personalized future blood glucose level prediction,” in *CEUR Workshop Proceedings*, vol. 2675, pp. 100–104, 2020.
- [53] M. Mayo and T. Koutny, “Neural multi-class classification approach to blood glucose level forecasting with prediction uncertainty visualisation,” in *KDH 2020*, vol. 2675, pp. 80–84, CEUR Workshop Proceedings, 2020.
- [54] G. Cappon, L. Meneghetti, F. Prendin, J. Pavan, G. Sparacino, S. Del Favero, A. Facchinetti, *et al.*, “A personalized and interpretable deep learning based approach to predict blood glucose concentration in type 1 diabetes,” in *KDH@ ECAI*, pp. 75–79, 2020.
- [55] KDH-2018, “The 3rd international workshop on knowledge discovery in healthcare data,” 2018. Retrieved from: <https://sites.google.com/view/kdhd-2018/>.
- [56] KDH-2020, “The 5rd international workshop on knowledge discovery in healthcare data,” 2020. Retrieved from: <https://sites.google.com/view/kdh-2020/bg1p-challenge>.
- [57] C. Marling and R. Bunescu, “The OhioT1DM dataset for blood glucose level prediction: Update 2020,” in *The 5th KDH workshop, ECAI 2020*, pp. 71–74, 2020.
- [58] J. Martinsson, A. Schliep, B. Eliasson, C. Meijner, S. Persson, and O. Mogren, “Automatic blood glucose prediction with confidence using recurrent neural networks,” in *3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ IJCAI-ECAI 2018, 13 July 2018*, pp. 64–68, 2018.

- [59] A. J. Graveling and B. M. Frier, “The risks of nocturnal hypoglycaemia in insulin-treated diabetes,” *Diabetes research and clinical practice*, vol. 133, pp. 30–39, 2017.
- [60] S. Pühr, M. Derdzinski, J. B. Welsh, A. S. Parker, T. Walker, and D. A. Price, “Real-world hypoglycemia avoidance with a continuous glucose monitoring system’s predictive low glucose alert,” *Diabetes technology & therapeutics*, vol. 21, no. 4, pp. 155–158, 2019.
- [61] E. R. Seaquist, J. Anderson, B. Childs, P. Cryer, S. Dagogo-Jack, L. Fish, S. R. Heller, H. Rodriguez, J. Rosenzweig, and R. Vigersky, “Hypoglycemia and diabetes: a report of a workgroup of the american diabetes association and the endocrine society,” *The Journal of Clinical Endocrinology & Metabolism*, vol. 98, no. 5, pp. 1845–1859, 2013.
- [62] J. Fleischer, T. K. Hansen, and S. L. Cichosz, “Hypoglycemia event prediction from cgm using ensemble learning,” *Frontiers in Clinical Diabetes and Healthcare*, vol. 3, p. 1066744, 2022.
- [63] S. L. Cichosz, J. Frystyk, O. K. Hejlesen, L. Tarnow, and J. Fleischer, “A novel algorithm for prediction and detection of hypoglycemia based on continuous glucose monitoring and heart rate variability in patients with type 1 diabetes,” *Journal of diabetes science and technology*, vol. 8, no. 4, pp. 731–737, 2014.
- [64] A. Z. Woldaregay, E. Årsand, S. Walderhaug, D. Albers, L. Mamykina, T. Botsis, and G. Hartvigsen, “Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes,” *Artificial intelligence in Medicine*, vol. 98, pp. 109–134, 2019.
- [65] L. Vu, S. Kefayati, T. Idé, V. Pavuluri, G. Jackson, L. Latts, Y. Zhong, P. Agrawal, and Y.-C. Chang, “Predicting nocturnal hypoglycemia from continuous glucose monitoring data with extended prediction horizon,” in *AMIA Annual Symposium Proceedings*, vol. 2019, p. 874, American Medical Informatics Association, 2019.
- [66] M. H. Jensen, C. Dethlefsen, P. Vestergaard, and O. Hejlesen, “Prediction of nocturnal hypoglycemia from continuous glucose monitoring data in people with type 1 diabetes:

- a proof-of-concept study,” *Journal of diabetes science and technology*, vol. 14, no. 2, pp. 250–256, 2020.
- [67] A. Bertachi, C. Viñals, L. Biagi, I. Contreras, J. Vehí, I. Conget, and M. Giménez, “Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor,” *Sensors*, vol. 20, no. 6, p. 1705, 2020.
- [68] A. Güemes, G. Cappon, B. Hernandez, M. Reddy, N. Oliver, P. Georgiou, and P. Herrero, “Predicting quality of overnight glycaemic control in type 1 diabetes using binary classifiers,” *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1439–1446, 2019.
- [69] C. Mosquera-Lopez, R. Dodier, N. S. Tyler, L. M. Wilson, J. El Youssef, J. R. Castle, and P. G. Jacobs, “Predicting and preventing nocturnal hypoglycemia in type 1 diabetes using big data analytics and decision theoretic analysis,” *Diabetes technology & therapeutics*, vol. 22, no. 11, pp. 801–811, 2020.
- [70] I. Afentakis, R. Unsworth, P. Herrero, N. Oliver, M. Reddy, and P. Georgiou, “Development and validation of binary classifiers to predict nocturnal hypoglycemia in adults with type 1 diabetes,” *Journal of Diabetes Science and Technology*, p. 19322968231185796, 2023.
- [71] J. Doupis, G. Festas, C. Tsilivigos, V. Efthymiou, and A. Kokkinos, “Smartphone-based technology in diabetes management,” *Diabetes Therapy*, vol. 11, no. 3, pp. 607–619, 2020.
- [72] D. Rhyner, H. Loher, J. Dehais, M. Anthimopoulos, S. Shevchik, R. H. Botwey, D. Duke, C. Stettler, P. Diem, and S. Mougiakakou, “Carbohydrate estimation by a mobile phone-based system versus self-estimations of individuals with type 1 diabetes mellitus: a comparative study,” *Journal of medical Internet research*, vol. 18, no. 5, p. e101, 2016.
- [73] Z. Huang, M. Soljak, B. O. Boehm, and J. Car, “Clinical relevance of smartphone apps for diabetes management: A global overview,” *Diabetes/metabolism research and reviews*, vol. 34, no. 4, p. e2990, 2018.

- [74] O. AlShorman, B. AlShorman, M. Alkhassaweneh, and F. Alkahtani, “A review of internet of medical things (iomt)-based remote health monitoring through wearable sensors: a case study for diabetic patients,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 1, pp. 414–422, 2020.
- [75] A. Chakrabarty, S. Zavitsanou, T. Sowrirajan, F. J. Doyle III, and E. Dassau, “Getting iot-ready: The face of next generation artificial pancreas systems,” in *The Artificial Pancreas*, pp. 29–57, Elsevier, 2019.
- [76] P. Herrero, M. El-Sharkawy, J. Daniels, N. Jugnee, C. N. Uduku, M. Reddy, N. Oliver, and P. Georgiou, “The bio-inspired artificial pancreas for type 1 diabetes control in the home: system architecture and preliminary results,” *Journal of diabetes science and technology*, vol. 13, no. 6, pp. 1017–1025, 2019.
- [77] P. Herrero, M. El Sharkawy, P. Pesl, M. Reddy, N. Oliver, D. Johnston, C. Toumazou, and P. Georgiou, “Live demonstration: A handheld bio-inspired artificial pancreas for treatment of diabetes,” in *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, pp. 172–172, IEEE, 2014.
- [78] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, “Dilated recurrent neural networks for glucose forecasting in type 1 diabetes,” *Journal of Healthcare Informatics Research*, vol. 4, no. 3, pp. 308–324, 2020.
- [79] T. Zhu, L. Kuang, K. Li, J. Zeng, P. Herrero, and P. Georgiou, “Blood glucose prediction in type 1 diabetes using deep learning on the edge,” *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2021.
- [80] M. M. H. Shuvo and S. K. Islam, “Deep multitask learning by stacked long short-term memory for predicting personalized blood glucose concentration,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1612–1623, 2023.
- [81] F. Prendin, S. Del Favero, M. Vettoretti, G. Sparacino, and A. Facchinetti, “Forecasting of glucose levels and hypoglycemic events: head-to-head comparison of linear and nonlin-

- ear data-driven algorithms based on continuous glucose monitoring data only,” *Sensors*, vol. 21, no. 5, p. 1647, 2021.
- [82] L. Bally, H. Thabit, S. Hartnell, E. Andereggen, Y. Ruan, M. E. Wilinska, M. L. Evans, M. M. Wertli, A. P. Coll, C. Stettler, *et al.*, “Closed-loop insulin delivery for glycemic control in noncritical care,” *New England Journal of Medicine*, vol. 379, no. 6, pp. 547–556, 2018.
- [83] A. Wallia, G. E. Umpierrez, R. J. Rushakoff, D. C. Klonoff, D. J. Rubin, S. Hill Golden, C. B. Cook, B. Thompson, and D. C. G. M. in the Hospital Panel, “Consensus statement on inpatient use of continuous glucose monitoring,” *Journal of Diabetes Science and Technology*, vol. 11, no. 5, pp. 1036–1044, 2017.
- [84] M. S. Hughes, A. Addala, and B. Buckingham, “Digital technology for diabetes,” *New England Journal of Medicine*, vol. 389, no. 22, pp. 2076–2086, 2023.
- [85] G. M. Davis, E. Faulds, T. Walker, D. Vigliotti, M. Rabinovich, J. Hester, L. Peng, B. McLean, P. Hannon, N. Poindexter, *et al.*, “Remote continuous glucose monitoring with a computerized insulin infusion protocol for critically ill patients in a COVID-19 medical ICU: proof of concept,” *Diabetes Care*, vol. 44, no. 4, pp. 1055–1058, 2021.
- [86] R. R. Longo, H. Elias, M. Khan, and J. J. Seley, “Use and accuracy of inpatient cgm during the covid-19 pandemic: an observational study of general medicine and icu patients,” *Journal of Diabetes Science and Technology*, vol. 16, no. 5, pp. 1136–1143, 2022.
- [87] D. C. Klonoff, K. T. Nguyen, N. Y. Xu, A. Gutierrez, J. C. Espinoza, and A. P. Vidmar, “Use of continuous glucose monitors by people without diabetes: an idea whose time has come?,” *Journal of Diabetes Science and Technology*, p. 19322968221110830, 2022.
- [88] S. Berry, K. Bermingham, A. Valdes, P. Franks, J. Wolf, and T. Spector, “Optimised glucose “time in range” using continuous glucose monitors in 4,805 non-diabetic individuals is associated with favourable diet and health: The ZOE PREDICT studies,” *Current Developments in Nutrition*, vol. 6, p. 1108, 2022.

- [89] L. PREECHASUK, S. K. RILSTONE, W. XI TANG, J. MAN, M. YANG, E. ZHAO, L. HOQUE, E. TUNCAY, P. WILDING, O. HALSE, *et al.*, “933-p: Glycemic level and glycemic variability in acute ischemic stroke and functional outcome at discharge—a continuous glucose monitoring study,” *Diabetes*, vol. 72, no. Supplement_1, 2023.
- [90] R. M. Bergenstal, R. W. Beck, K. L. Close, G. Grunberger, D. B. Sacks, A. Kowalski, A. S. Brown, L. Heinemann, G. Aleppo, D. B. Ryan, *et al.*, “Glucose management indicator (GMI): a new term for estimating a1c from continuous glucose monitoring,” *Diabetes care*, vol. 41, no. 11, pp. 2275–2280, 2018.
- [91] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.
- [92] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [93] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with lstm,” *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [94] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, “Evaluating clinical accuracy of systems for self-monitoring of blood glucose,” *Diabetes Care*, vol. 10, no. 5, pp. 622–628, 1987.
- [95] J. L. Parkes, S. L. Slatin, S. Pardo, and B. H. Ginsberg, “A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose,” *Diabetes care*, vol. 23, no. 8, pp. 1143–1148, 2000.
- [96] A. Pfützner, D. C. Klonoff, S. Pardo, and J. L. Parkes, “Technical aspects of the parkes error grid,” *Journal of Diabetes Science and Technology*, vol. 7, no. 5, pp. 1275–1281, 2013.
- [97] T. Zhu, L. Kuang, J. Daniels, P. Herrero, K. Li, and P. Georgiou, “IoMT-enabled real-time blood glucose prediction with deep learning and edge computing,” *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3706–3719, 2023.

- [98] A. Ceriello, L. Monnier, and D. Owens, “Glycaemic variability in diabetes: clinical and therapeutic implications,” *The Lancet Diabetes & Endocrinology*, vol. 7, no. 3, pp. 221–230, 2019.
- [99] M. I. Maiorino, G. Bellastella, O. Casciano, P. Cirillo, V. Simeon, P. Chiodini, M. Petrizzo, M. Gicchino, O. Romano, P. Caruso, *et al.*, “The effects of subcutaneous insulin infusion versus multiple insulin injections on glucose variability in young adults with type 1 diabetes: the 2-year follow-up of the observational metro study,” *Diabetes Technology & Therapeutics*, vol. 20, no. 2, pp. 117–126, 2018.
- [100] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [101] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning,” *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 193–204, 2023.
- [102] R. Unsworth, R. Armiger, N. Jugnee, M. Thomas, P. Herrero, P. Georgiou, N. Oliver, and M. Reddy, “Safety and efficacy of an adaptive bolus calculator for type 1 diabetes: A randomized controlled crossover study,” *Diabetes Technology & Therapeutics*, vol. 25, no. 6, pp. 414–425, 2023.
- [103] M. Reddy, N. Jugnee, A. El Laboudi, E. Spanudakis, S. Anantharaja, and N. Oliver, “A randomized controlled pilot study of continuous glucose monitoring and flash glucose monitoring in people with type 1 diabetes and impaired awareness of hypoglycaemia,” *Diabetic Medicine*, vol. 35, no. 4, pp. 483–490, 2018.
- [104] M. Reddy, N. Jugnee, S. Anantharaja, and N. Oliver, “Switching from flash glucose monitoring to continuous glucose monitoring on hypoglycemia in adults with type 1 diabetes at high hypoglycemia risk: the extension phase of the i hart cgm study,” *Diabetes technology & therapeutics*, vol. 20, no. 11, pp. 751–757, 2018.
- [105] L. Breiman, “Random forests,” *Machine learning*, vol. 45, pp. 5–32, 2001.

- [106] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*, pp. 1–15, Springer, 2000.
- [107] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [108] S. I. Gallant *et al.*, “Perceptron-based learning algorithms,” *IEEE Transactions on neural networks*, vol. 1, no. 2, pp. 179–191, 1990.
- [109] S. Raschka and V. Mirjalili, *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd, 2019.
- [110] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, “Tsfel: Time series feature extraction library,” *SoftwareX*, vol. 11, p. 100456, 2020.
- [111] A. L. McCall, D. J. Cox, J. Crean, M. Gloster, and B. P. Kovatchev, “A novel analytical method for assessing glucose variability: using cgms in type 1 diabetes mellitus,” *Diabetes technology & therapeutics*, vol. 8, no. 6, pp. 644–653, 2006.
- [112] M. E. Wilinska, L. J. Chassin, H. C. Schaller, L. Schaupp, T. R. Pieber, and R. Hovorka, “Insulin kinetics in type-1 diabetes: continuous and bolus delivery of rapid acting insulin,” *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 1, pp. 3–12, 2004.
- [113] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, “Comparative study of techniques for large-scale feature selection,” in *Machine intelligence and pattern recognition*, vol. 16, pp. 403–413, Elsevier, 1994.
- [114] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [115] M. C. Riddell, I. W. Gallen, C. E. Smart, C. E. Taplin, P. Adolfsson, A. N. Lumb, A. Kowalski, R. Rabasa-Lhoret, R. J. McCrimmon, C. Hume, *et al.*, “Exercise manage-

- ment in type 1 diabetes: a consensus statement,” *The lancet Diabetes & endocrinology*, vol. 5, no. 5, pp. 377–390, 2017.
- [116] N. S. Tyler, C. M. Mosquera-Lopez, L. M. Wilson, R. H. Dodier, D. L. Branigan, V. B. Gabo, F. H. Guillot, W. W. Hiltz, J. El Youssef, J. R. Castle, *et al.*, “An artificial intelligence decision support system for the management of type 1 diabetes,” *Nature metabolism*, vol. 2, no. 7, pp. 612–619, 2020.
- [117] S. G. Mougiakakou, C. S. Bartsocas, E. Bozas, N. Chaniotakis, D. Iliopoulou, I. Kouris, S. Pavlopoulos, A. Prountzou, M. Skevofilakas, A. Tsoukalis, *et al.*, “Smartdiab: a communication and information technology approach for the intelligent monitoring, management and follow-up of type 1 diabetes patients,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 622–633, 2010.
- [118] C. Liu, P. Avari, Y. Leal, M. Wos, K. Sivasithamparam, P. Georgiou, M. Reddy, J. M. Fernández-Real, C. Martin, M. Fernández-Balsells, *et al.*, “A modular safety system for an insulin dose recommender: a feasibility study,” *Journal of diabetes science and technology*, vol. 14, no. 1, pp. 87–96, 2020.
- [119] A. Hásková, L. Radovnická, L. Petruželková, C. G. Parkin, G. Grunberger, E. Horová, V. Navrátilová, O. Kádě, M. Matoulek, M. Prázný, *et al.*, “Real-time cgm is superior to flash glucose monitoring for glucose control in type 1 diabetes: the corrida randomized controlled trial,” *Diabetes Care*, vol. 43, no. 11, pp. 2744–2750, 2020.
- [120] M. M. Visser, S. Charleer, S. Fieuws, C. De Block, R. Hilbrands, L. Van Huffel, T. Maes, G. Vanhaverbeke, E. Dirinck, N. Myngher, *et al.*, “Comparing real-time and intermittently scanned continuous glucose monitoring in adults with type 1 diabetes (alertt1): a 6-month, prospective, multicentre, randomised controlled trial,” *The Lancet*, vol. 397, no. 10291, pp. 2275–2283, 2021.
- [121] B. P. Kovatchev, M. Breton, C. Dalla Man, and C. Cobelli, “In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes,” 2009.

- [122] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, “The uva/padova type 1 diabetes simulator: new features,” *Journal of diabetes science and technology*, vol. 8, no. 1, pp. 26–34, 2014.
- [123] R. Visentin, E. Campos-Náñez, M. Schiavon, D. Lv, M. Vettoretti, M. Breton, B. P. Kovatchev, C. Dalla Man, and C. Cobelli, “The uva/padova type 1 diabetes simulator goes from single meal to single day,” *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 273–281, 2018.
- [124] L. Magni, D. M. Raimondo, C. D. Man, M. Breton, S. Patek, G. De Nicolao, C. Cobelli, and B. P. Kovatchev, “Evaluating the efficacy of closed-loop glucose regulation via control-variability grid analysis,” *Journal of diabetes science and technology*, vol. 2, no. 4, pp. 630–635, 2008.
- [125] P. Herrero, P. Pesl, J. Bondia, M. Reddy, N. Oliver, P. Georgiou, and C. Toumazou, “Method for automatic adjustment of an insulin bolus calculator: in silico robustness evaluation under intra-day variability,” *Computer methods and programs in biomedicine*, vol. 119, no. 1, pp. 1–8, 2015.
- [126] A. Brazeau, H. Mircescu, K. Desjardins, C. Leroux, I. Strychar, J. Ekoé, and R. Rabasa-Lhoret, “Carbohydrate counting accuracy and blood glucose variability in adults with type 1 diabetes,” *Diabetes research and clinical practice*, vol. 99, no. 1, pp. 19–23, 2013.
- [127] A. Parcerisas, I. Contreras, A. Delecourt, A. Bertachi, A. Beneyto, I. Conget, C. Viñals, M. Giménez, and J. Vehi, “A machine learning approach to minimize nocturnal hypoglycemic events in type 1 diabetic patients under multiple doses of insulin,” *Sensors*, vol. 22, no. 4, p. 1665, 2022.
- [128] D. Care *et al.*, “Standards of care in diabetes—2023,” *Diabetes care*, vol. 46, pp. S1–S267, 2023.
- [129] N. Diabetes and J. B. D. Societies, “The hospital management of hypoglycaemia in adults with diabetes mellitus,” Available at www.diabetologists-abcd.org.uk/JBDS/JBDS_IP_Hypo_Adults.pdf (accessed March 2015), 2010.

- [130] M. Kalergis, A. Schiffrin, R. Gougeon, P. J. Jones, and J.-F. Yale, “Impact of bedtime snack composition on prevention of nocturnal hypoglycemia in adults with type 1 diabetes undergoing intensive insulin management using lispro insulin before meals: a randomized, placebo-controlled, crossover trial,” *Diabetes Care*, vol. 26, no. 1, pp. 9–15, 2003.
- [131] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.