The University Of Sheffield.

Artificial Intelligence in Blood Glucose Level Prediction for Type 1 Diabetes Management

by

Hoda Nemat

Dissertation submitted to the University of Sheffield in partial
fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Electronic and Electrical Engineering

August 2023

Acknowledgments

I would like to express my gratitude to my supervisors Prof. Mohammed Benaissa, and Dr. Jackie Elliott who have guided and supported me throughout my PhD. I am grateful to my friend and colleague, Mr. Heydar Khadem, for his insightful comments. I also, would like to thank Dr. Mohammed R. Eiassa, for the discussions we had at our group meetings.

I dedicate this thesis to my family and thank them for their support.

Contents

List of Figures

List of Tables

Acronyms

ADF Augmented Dickey-Fuller
AI Artificial Intelligence
ANN Artificial Neural Network
AP Artificial Pancreas
AR Autoregression
ARIMA Autoregressive Integrated Moving Average
ARIMAX Autoregressive Integrated Moving Average with Exogenous Variables
ARMA Autoregressive Moving Average
ARMAX Autoregressive Moving Average with Exogenous Variables
ARX Autoregression with Exogenous Variables
BGL Blood Glucose Level
BiLSTM Bidirectional Long Short-Term Memory
CCM Convergent Cross Mapping
CD Critical Difference
CGM Continues Glucose Monitoring
CNN Convolution Neural Networks
ConvLSTM Convolutional Long Short-Term Memory
CTF Classical Time series Forecasting
DNN Deep Neural Network
ECCM Extended Convergent Cross Mapping
FNN Feed-forward Neural Network
GDM Gestational Diabetes Mellitus
GRU Gated Recurrent Unit
GSR Galvanic Skin Response
HbA1c Haemoglobin A1c
HR Heart Rate
ISF Interstitial Fluid
KDD Knowledge Discovery in Databases
KPSS Kwiatkowski–Phillips–Schmidt–Shin

LSTM Long Short-Term Memory
MA Magnitude of Acceleration
MAE Mean Absolute Error
MCC Matthews correlation coefficient
MLP Multilayer Perceptron
MSE Mean Square Error
PA Physical Activity
PLSR Partial Least Squares Regression
Relu Rectified Linear Unit
RMSE Root Mean Square Error
RNN Recurrent Neural Network
SC Step Count
SE surveillance error
SEG surveillance error grid
SMBG Self-Monitoring of Blood Glucose
ST Skin Temperature
SVM Support Vector Machine
SVR Support Vector Regression
T1DM Type 1 Diabetes Mellitus
T2DM Type 2 Diabetes Mellitus
TML Traditional Machine Learning
VLSTM Vanilla Long Short-Term Memory

Abstract

Effective management of type 1 diabetes mellitus (T1DM) reduces the associated complications. T1DM management aims to maintain blood glucose levels (BGLs) within a target range. BGL prediction is an important tool to help maximise the time BGL is in the target range and thus minimise both acute and chronic diabetes-related complications. Data-driven BGL prediction models estimate future BGL utilising current and past information and provide early warnings concerning inadequate glycaemic control. Despite many works performed on BGL prediction, further improvements in prediction accuracy are still desired. This thesis focuses on BGL prediction in T1DM using artificial intelligence.

As part of this thesis, advanced artificial intelligence techniques, including deep learning, ensemble learning, causal analysis, and data fusion are explored to enhance the performance of BGL prediction. Leveraging deep learning and ensemble learning, three deep-ensemble models are proposed. The superior performance of the proposed ensemble models over non-ensemble benchmark models is shown. Also, the relations between BGL and affecting variables, including carbohydrate intake, injected bolus insulin, and physical activity, via the causality context are examined. Then, by proposing novel approaches, leveraging causality information as prior knowledge for BGL prediction is investigated. The results show the effectiveness of using causality information in BGL prediction. Moreover, new approaches for extracting information from physical activity, as a crucial factor in T1DM management, are developed, and the fusion of this information with BGL data at multiple levels is explored. Based on the results, incorporating physical activity into BGL prediction can improve prediction performance. Furthermore, the performance of different data-driven time series forecasting approaches with different inputs for BGL prediction is examined and assessed to provide useful information regarding the primary choices of the model structure and input.

Chapter 1

Introduction

1.1 Diabetes overview

Diabetes mellitus, also known as diabetes, is a growing metabolic disorder and a significant cause of mortality worldwide that can cause severe complications. The disease has massive economic implications related to the direct costs of disease treatment and the indirect costs relevant to mortality and morbidity. Diabetes is characterised by the lack of insulin secretion from the pancreas, insulin sensitivity of body cells, or a combination of both. There are genetic and environmental risk factors that contribute to the development of the disease. A glycated haemoglobin A1c (HbA1c) test, which is a measure of average blood glucose level (BGL) over the course of the past two to three months, can be used to diagnose diabetes in clinic [1, 2].

1.1.1 Types of diabetes

The main types of diabetes are generally categorised as type 1 diabetes mellitus (T1DM), type 2 diabetes mellitus (T2DM), and gestational diabetes mellitus (GDM). T1DM is characterised by an absolute insulin deficiency, caused by the destruction of pancreatic beta cells, responsible for the production of insulin. Hence, people with T1DM require insulin therapy for the management of their disease. T2DM is characterised by insulin resistance. In T2DM, the pancreas is capable of producing insulin, however, the body cells become resistant to the effect of insulin. Also, over time the pancreas may lose its ability to secrete enough insulin, so there is a relative deficiency. The medications that are given to T2DM patients improve the secretion or absorption of insulin. After many years of T2DM, some patients secrete so little insulin that they also need insulin therapy. In GDM, the interac-

tion between insulin and pregnancy-related hormones results in insufficient insulin to meet the extra demands of insulin resistance during pregnancy. This condition usually resolves after pregnancy [2, 3].

1.1.2 Glycaemic control

The key role of diabetes management is glycaemic control which is controlling BGL to maintain the normal range. Unlike healthy individuals whose BGLs remain in the normal range, people with T1DM can experience high or low BGLs. Blood glucose concentration is normally reported in milligrams per decilitre (mg/dl) or millimoles per litre (mmol/l) units. Based on the value of blood glucose concentration, three glycaemic states are defined; hypoglycaemia, normoglycaemia, and hyperglycaemia. BGL normally should be in the range of 70 mg/dl and 180 mg/dl, which is also called normoglycaemia. Hypoglycaemia is a situation in which BGL is below 70 mg/dl. Hypoglycaemia can be classified into three levels: mild (54 mg/dl < BGL < 70 mg/dl), moderate (40 mg/dl < BGL < 54 mg/dl), and severe (BGL < 40 mg/dl). hyperglycaemia refers to a condition in which BGL is above 180 mg/dl. hyperglycaemia can also be classified into three levels: mild (180 mg/dl < BGL < 250 mg/dl), moderate (250 mg/dl < BGL < 320 mg/dl), and severe (BGL > 320 mg/dl). The main limiting factor during the control of glycaemia is hypoglycaemia, which is associated with acute increased morbidity and mortality. So the primary challenge in diabetes management is the correction of hyperglycaemia without hypoglycaemia occurrence. To cope with this challenge, people with diabetes need to monitor their BGL frequently, and adjust medications accordingly [4, 5].

1.1.3 Glycaemic monitoring

There are three main glycaemic monitoring methods; HbA1c test, as a long-term metric, self-monitoring of blood glucose (SMBG) and continuous glucose monitoring (CGM) sensors as short-term metrics [6]. The HbA1c test is a laboratory test which determines the percentage of HbA1c in the blood. This test provides an indicator of the average BGL during the past two to three months. It can be presented either in percentage (%) or millimoles per mole (mmol/mol) units. HbA1c is accepted as a gold standard marker for average glycaemic control, however, it does not measure glycaemic variability [2].

SMBG was the first capillary measurement of BGL by providing a snapshot of the BGL at the time of measurement. Using glucometers, patients with diabetes

need to prick their fingers to collect a blood sample. They then apply the sample on a test strip which is connected to a pocket-sized device. The device then identifies the glucose concentration using electrochemical, colourimetric or optical procedures. SMBG can be used at all times of the day. Most patients are advised to perform SMBG four to eight times a day specially pre-meals, occasionally after-meals, bedtime, pre-exercise, and when there is a suspicion of hypoglycaemia. There is a positive association between the frequency of SMBG and glycaemic control improvement. This glycaemic monitoring method is comparably inexpensive and easy to learn, however, it is inconvenient and painful. Moreover, each measurement presents only a snapshot of blood glucose levels, so it may miss glucose excursions [7].

To overcome the limitations of the SMBG method, CGM sensors have been introduced. They are portable devices that measure glucose concentration regularly and provide comprehensive glycaemic monitoring. Based on the fact that the glucose concentration of interstitial fluid (ISF) is similar to blood glucose concentration, CGM devices measure ISF glucose concentration. For this purpose, an electrode placed under the skin senses the glucose in the ISF and sends the signal wirelessly to an external apparatus which identifies the BGL. CGM sensors can continuously monitor BGL and assist people with diabetes to make more precise decisions about their glycaemic control. CGM also provides a variety of glycaemic factors such as time in glycaemic target, time in hypoglycaemia, glucose excursion, and intra- and inter-day glucose variability [7, 8].

1.1.4 Diabetes complications

Deviations from normal BGL in diabetes can result in different short-term and long-term complications. The main short-term diabetic complications include hypoglycaemia and hyperglycaemia which refer to the occurrence of low and high BGL, respectively. The main long-term diabetic complications are divided into microvascular and macrovascular complications. Microvascular complications are characterised by damage to small blood vessels, such as diabetic neuropathy, nephropathy, and retinopathy, leading to amputations, kidney failure and blindness respectively. Whereas macrovascular complications affect large blood vessels, such as coronary artery disease, peripheral artery disease and strokes. The occurrence of these complications can be delayed or even prevented by effective management of the disease [9].

1.2 Type 1 diabetes management

The literature has emphasised the importance of the self-management of T1DM, especially in reducing complications associated with the disease. In T1DM, insulin therapy is required for the management of the disease and control of glycaemia. In T1DM management, glycaemic control can be affected by a number of factors. The main affecting factors include carbohydrate intake, physical activity (PA), and insulin dose.

1.2.1 Carbohydrate intake

Meals greatly impact glycaemic control in T1DM patients. Food consumption affects BGL in a number of ways due to different physiological effects. In healthy individuals, absorbing carbohydrates from foods results in a temporary increase in BGL. The increase is automatically detected by the pancreas, and insulin secretion starts to return the BGL to a fasting level. The challenge of food consumption for people with T1DM is how BGL can be lowered in the shortest time after glucose absorption from carbohydrates. Fats, fibres, and proteins in foods can delay, slow, or decrease the glucose absorption process. Several measures have been introduced to quantify the impact of food on BGL, including the glycaemic index and the glycaemic load. These measures should be considered by patients for their glycaemic control [5].

1.2.2 Physical activity

Regular PA can improve insulin sensitivity and reduce the risk of cardiovascular disease. Depending on the type, form, intensity, and duration of exercise, PA can significantly affect BGL in patients with T1DM [10, 11]. However, proper management of glycaemia during and after PA is challenging for both patients and clinicians. It is difficult to accurately adjust insulin and carbohydrate during and after PA, and any mistake may cause hypoglycaemia or hyperglycaemia [12].

1.2.3 Insulin therapy

Individuals with T1DM depend on injected external insulin to compensate for the lack of insulin secretion in the body. In general, two types of insulin are used: basal insulin and bolus insulin. Basal, as the background insulin, controls BGL in between meals, whereas bolus is used to manage BGL with meals. T1DM patients are required to consider the effect of their carbohydrate intake and PA levels when

adjusting bolus doses to control their BGL, aiming for the normoglycaemic range [13, 14].

Various insulin therapy solutions are available, including multiple daily injections, insulin pumps, and artificial pancreas (AP) systems. In multiple daily injections, patients with T1DM regulate their BGL by administering several injections of bolus and basal, daily. Insulin pumps, which administer insulin via an infusion cannula, are open-loop systems that require patients to adjust insulin dosage manually. An AP is a closed-loop glucose control system that mimics the function of a healthy pancreas to regulate BGL for T1DM patients. An AP system typically consists of a CGM sensor, an insulin pump, and a controlling algorithm to adjust the insulin dose based on the CGM information. The primary objective of an AP system is to determine the optimal insulin dose to maintain BGL in the normal range and to avoid occurrences of adverse glycaemic events including hypoglycaemia or hyperglycaemia [5]. However, due to the subcutaneous nature of insulin delivery, current systems have a significant time lag compared to healthy individuals, and the amount of carbohydrates consumed needs to be manually announced.

1.3 Artificial intelligence

In general, intelligence is described as a set of capabilities, including analysing, learning, and reasoning that can be used for solving problems and making decisions. Artificial intelligence (AI) is a field of computer science that aims to program computers to simulate human intelligence in order to analyse information and make sophisticated inferences [15, 16]. Main AI approaches can be categorised into three groups based on their objectives; learning from knowledge, reasoning from knowledge, and discovery of knowledge. The primary objective of learning in AI is to enable computers to acquire knowledge on their own without the intervention of humans. The most significant strategies for learning from knowledge are artificial neural networks (ANNs) and support vector machines (SVMs). In the context of reasoning from knowledge, it refers to the development of more accurate and robust methods for drawing inferences from knowledge and making conclusions. In knowledge discovery, also called knowledge discovery in databases (KDD), algorithms are developed for the retrieval of potential information from databases [17].

KDD includes theories, methods and techniques, attempting to extract valid and understandable knowledge from data, and is associated with the assessment and explanation of patterns. Hence, it needs a huge knowledge about the context of the study. KDD has different steps including data selection, preprocessing, transfor-

mation, data mining, and interpretation/evaluation. The most significant step in the KDD is data mining which involves the applicability of machine learning algorithms in data analysis [1].

1.3.1 Machine learning

Machine learning is a computer program that learns and improves from experience. Machine learning tasks with respect to the learning process can be categorised as supervised, unsupervised, or reinforcement learning. Supervised learning is related to inducing a function from labelled data, while unsupervised learning is related to inducing structures of unlabeled data. In supervised learning, the system learns a target function as a representation of a model describing the data. The objective function, then predicts the output variable from the input variables. Based on the type of output which is discrete classes or continuous values, supervised learning tasks are divided into classification or regression, respectively [18, 19].

1.3.2 Time series forecasting

An important aspect of AI in analysing and predicting data with applications in various areas, including climate, finance, and medicine, is time series forecasting. Time series forecasting, using historical data, attempts to find underlying patterns in the data to anticipate future time-dependent events. According to the number of variables measured over time, a time series forecasting can be univariate or multivariate. In univariate forecasting, past measurements or observations of a single variable are used to predict future values, while in multivariate forecasting, multiple time series are employed to make predictions [20, 21, 22].

1.4 Artificial intelligence in diabetes

AI applications, along with medical devices and sensor technologies, have hugely impacted healthcare and enabled potentially better diagnosis and management of chronic diseases like diabetes [23, 24, 25]. In the field of diabetes care, patients, clinicians, and healthcare systems can benefit from AI. By providing personalised assistance, AI enables patients to be informed, have continuous monitoring, and be able to independently manage their diabetes. Additionally, healthcare professionals can utilise AI for decision-making support. Through numerous contributions to clinical decision support, risk stratification prediction, and patient self-management

tools, AI has the potential to revolutionise the prevention, diagnosis, and management of diabetes. More specifically, applications of AI in diabetes management include prediction and detection of adverse glycaemic events, advisory systems for calculating insulin bolus, and BGL prediction. In this regard, BGL prediction is one of the most widely used AI applications in diabetes management [15, 17, 26, 27].

1.5 Blood glucose level prediction

Advancements in medical sensors have enabled automatic continuous personal data collection. Furthermore, the developments of mobile health applications utilising AI strategies have advanced the self-monitoring and management of diabetes. One of the most important diabetes-related applications is the BGL alarm which is based on an accurate BGL prediction model. It is proven that predicting BGL is a promising tool for glycaemic control. A BGL alarm significantly contributes to glycaemic control by providing patients with risk alerts, which allow them to take corrective precautionary actions based on the predicted BGL. This involves taking glucose prompted by a hypoglycaemia alarm and injecting correction insulin boluses following a hyperglycaemia alarm.

Furthermore, estimating BGL in a given prediction horizon in advance is the most significant feature of closed-loop AP systems, as the most advanced solution for T1DM management, to prevent adverse glycaemic events. Hence, obtaining more reliable models for predicting BGL is a critical AI-related component of these systems [5, 28, 29]. Consequently, any improvement in BGL prediction can advance diabetes management in both open-looped and closed-looped glycaemic control systems. Figure 1.1 illustrates the impact of BGL prediction in both glycaemic control systems in T1DM management.

In spite of this, accurate BGL prediction remains a challenge due to the influence of numerous factors, such as carbohydrates, bolus, and PA. In particular, among these influencing factors, PA presents a distinct challenge for BGL prediction. Part of the complexity of PA arises from the fact that its effect on BGL varies significantly from day to day, even for the same type and duration of the exercise performed at the same time and after similar meals [30].

1.6 Thesis scope

This thesis generally aims to contribute to T1DM management, and more specifically in BGL prediction, using AI methods. Although many studies have been

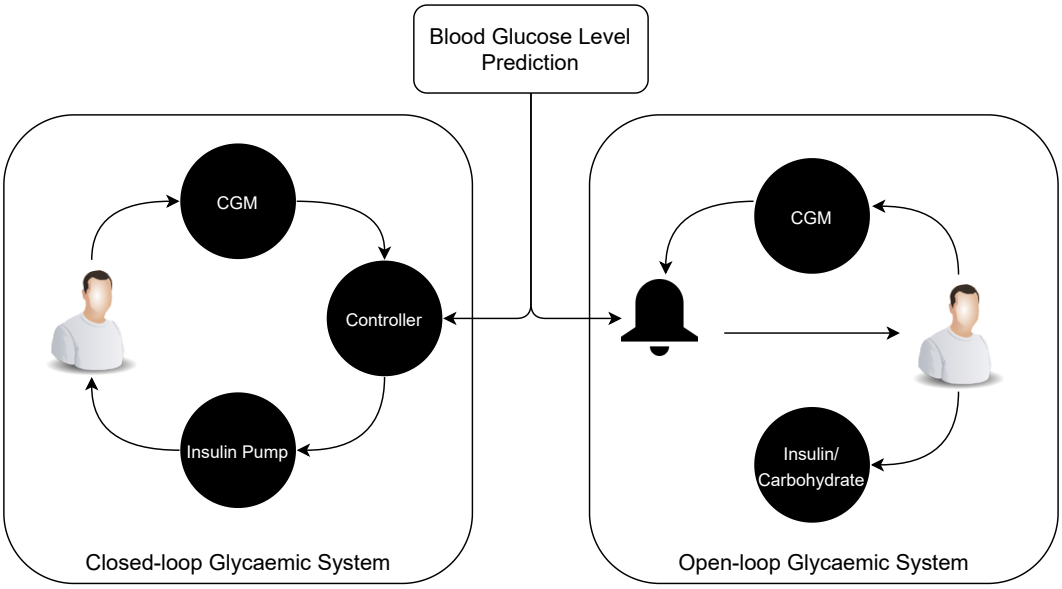


Figure 1.1: A schematic diagram illustrating the impact of blood glucose level prediction in both closed-loop and open-loop glycaemic systems on type 1 diabetes management.

conducted in this area and many achievements have been made, various challenges still remain to be addressed through further research. The specific challenges, objectives, and contributions for addressing them performed in the course of this thesis are outlined in this section.

1.6.1 Challenges

This thesis contributes to four of the numerous challenges in the field of BGL prediction. These specific challenges are described in the following.

Exploring advanced AI strategies in the field of BGL prediction, as a significant AI-based application in diabetes management, would be demanding given the continued growth and development of AI. Among different developments, ensemble learning and causality analysis are two advanced AI strategies that have recently emerged.

Ensemble learning, which involves leveraging multiple models for learning, has recently shown competitive performance in different areas including BGL prediction [31, 32]. Due to the complexity of glycaemic dynamics, a single model may not always be able to adapt to changes and accurately predict BGL in all hypoglycaemia, normoglycaemia, and hyperglycaemia events. While a model may provide a better prediction of BGL in one of the glycaemic events, it may have weak prediction power in another glycaemic event, where another model can provide a better

prediction. Accordingly, ensemble techniques are capable of improving the accuracy of BGL prediction by combining multiple models [19]. Hence, exploring the capabilities of new ensemble models in BGL prediction is beneficial and can help achieve more reliable predictions.

Causality analysis and, more specifically, causal inference, the study of the cause-and-effect relationship, is yet another AI advancement that has made achievements in time series forecasting in various fields [33, 34, 35]. As previously mentioned, carbohydrates, bolus, and PA have been identified as the main factors affecting the BGL [14]. Since there are cause and effect relations between these variables, BGL prediction may benefit from the investigation of these relations in a causal context. Literature on T1DM management lacks causal examination of affecting variables and causality-based BGL prediction approaches. Hence, developing effective methodologies for leveraging causal inference in BGL prediction is another field in which AI can contribute to T1DM management.

Another significant research area contributing to T1DM management is PA. Regular PA can help reduce the risk of cardiovascular disease, and other health conditions. However, due to the lack of explicit knowledge regarding how exactly PA impacts BGL, optimal diabetes management is hindered in the presence of PA. It is imperative to effectively deploy PA in BGL prediction to support open-loop and closed-loop diabetes management systems with the incorporation of this crucial factor [10, 12, 36, 37]. Although several works were performed considering PA in the BGL prediction, there is still a demand to discover optimal approaches for PA fusion with BGL in order to improve the performance of BGL prediction. Accordingly, it is beneficial to perform a rigorous investigation into extracting information from PA data and fusing data at different levels, including signal-level, feature-level, and decision-level fusion, to find more efficient ways of fusing PA and BGL data.

The choice of inputs, along with the fundamental choice of model structure, is another challenge in BGL prediction. Comparing the efficacy of different prediction models would be beneficial in the advancement of BGL prediction performance. However, using different datasets, different inputs, and different model settings has made this comparison difficult and limited studies [38, 39] were performed in this regard. Also, investigating to what extent other relevant variables can contribute to the performance of BGL prediction in different time series forecasting approaches would be another helpful factor in the advancement of BGL prediction. Limited work [40, 41] has been made to compare different inputs, which has resulted in different conclusions. Hence, a comprehensive investigation of different data-driven

time series forecasting approaches using different inputs along with rigorous statistical analyses and evaluation, can provide insightful findings in the context of BGL prediction.

1.6.2 Objectives and contributions

Four objectives have been devised to contribute to the above-mentioned challenges. In brief, the objectives of this thesis are:

- To develop novel advanced architectures for leveraging ensemble learning to enhance BGL prediction performance.
- To investigate the feasibility of leveraging causal analysis to enhance BGL prediction performance by developing novel causality-based prediction approaches.
- To develop novel approaches for extracting PA information and integrating this information at different levels in order to develop various PA-informed models for BGL prediction.
- To investigate and compare the efficacy of various data-driven time series forecasting approaches with different inputs for BGL prediction.

In order to accomplish these objectives, four contributions are made, which are briefly explained in the following. In the first contribution, we propose novel advanced architectures to predict BGL in people with T1DM using deep learning and ensemble learning. Two types of long short-term memory (LSTM) networks, including vanilla LSTM and bidirectional LSTM, along with a linear regression model, are used as base-learners in the ensemble architectures. These base models are also considered as benchmark BGL prediction models. In the advanced architectures, three meta-learning approaches are developed, two of which are novel. The performance of developed ensemble methods is compared with benchmark non-ensemble models and validated through evaluation and statistical analyses.

In the second contribution, the feasibility of using causality information as prior knowledge to improve BGL prediction performance is investigated. Initially, the relations between BGL and carbohydrates, bolus, and PA are investigated in the causality context. To accomplish this, the causal strengths of each variable with BGL are quantified using the convergent cross mapping method. Moreover, the optimal time lag for each variable is determined by utilising the extended convergent

cross mapping. Then, two novel approaches for leveraging quantified causality information in BGL prediction are proposed. In the first approach, causality strengths are used as weights for affecting input variables. In the second approach, the optimal causal lags and the corresponding causality strengths are considered shifts and weights for the input variables, respectively. Finally, to validate the impact of causality usage on BGL prediction performance, evaluation and statistical analyses are used to assess and compare the performance of BGL prediction with and without deploying causality.

In the third contribution, several novel PA-informed prediction models are developed by extracting and fusing PA-related information with BGL data at multiple levels, including signal-level, feature-level, and decision-level. For signal-level data fusion, fusing combinations of raw PA data directly collected from wristbands are examined. Also, for feature-level data fusion, three feature engineering approaches are developed; subjective assessments of PA, objective assessments of PA, and statistics of PA. Furthermore, in decision-level data fusion, ensemble learning is used to combine predictions from models trained with different inputs. The effectiveness of proposed prediction methods incorporating PA is then assessed using evaluation metrics and statistical analyses.

Finally, in the fourth contribution, we comprehensively investigate different data-driven BGL prediction approaches including classical time series forecasting, traditional machine learning, and deep neural networks. Also, a comparison between univariate input (BGL data only) and multivariate input (BGL data along with carbohydrate, bolus, and PA data) is performed to investigate how adding exogenous variables impacts different prediction approaches in BGL prediction. Rigorous evaluation and statistical analyses are then applied to compare the performance of different models and inputs.

1.7 Publications

The journal articles and conference papers that have been published or submitted throughout this thesis are presented in this section.

1.7.1 First-authored publications

1.7.1.1 Journal articles

- **H. Nemat**, H. Khadem, J. Elliott, and M. Benaissa, "Physical Activity Integration in Blood Glucose Level Prediction: Different Levels of Data Fusion,"

IEEE Journal of Biomedical and Health Informatics, [*Submitted*].

- **H. Nemat**, H. Khadem, J. Elliott, and M. Benaissa, "Data-driven Blood Glucose Level Prediction in Type 1 Diabetes: A Comprehensive Comparative Analysis," Scientific Reports, [*Under review*].

- **H. Nemat**, H. Khadem, J. Elliott, and M. Benaissa, "Causality analysis in type 1 diabetes mellitus with application to blood glucose level prediction," Computers in Biology and Medicine, p. 106535, 2023.

- **H. Nemat**, H. Khadem, M. R. Eissa, J. Elliott, and M. Benaissa, "Blood glucose level prediction: Advanced deep-ensemble learning approach," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 6, pp. 2758–2769, 2022.

1.7.1.2 Conference paper

- **H. Nemat**, H. Khadem, J. Elliott, and M. Benaissa, "Data fusion of activity and CGM for predicting blood glucose levels," in 5th International Workshop on Knowledge Discovery in Healthcare Data, vol. 2675, 2020, pp. 120–124.

1.7.2 Co-authored publications

1.7.2.1 Journal articles

- H. Khadem, **H. Nemat**, J. Elliott, and M. Benaissa, "Glucose Quantification from Absorption Spectroscopy: Benchmark of Machine Learning and Filtering Chemometric Techniques," Microchemical Journal, [*submitted*].

- H. Khadem, **H. Nemat**, J. Elliott, and M. Benaissa, "New Advanced Interdependent System Topologies for Deep Learning Nonlinear Time Series Forecasting," Journal of Neural Networks, [*Revised version submitted*].

- H. Khadem, **H. Nemat**, J. Elliott, and M. Benaissa, "Blood glucose level time series forecasting: Nested deep ensemble learning lag fusion," Bioengineering, vol. 10, no. 4, p. 487, 2023.

- H. Khadem, **H. Nemat**, J. Elliott, and M. Benaissa, "Interpretable machine learning for inpatient covid-19 mortality risk assessments: Diabetes mellitus exclusive interplay," Sensors, vol. 22, no. 22, p. 8757, 2022.

- H. Khadem, **H. Nemat**, M. R. Eissa, J. Elliott, and M. Benaissa, "Covid-19 mortality risk assessments for individuals with and without diabetes mellitus: Machine learning models integrated with interpretation framework," Computers in Biology and Medicine, vol. 144, p. 105361, 2022.

- H. Khadem, **H. Nemat**, J. Elliott, and M. Benaissa, "Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy," Talanta, vol. 243, p. 123379, 2022.

- H. Khadem, M. R. Eissa, **H. Nemat**, O. Alrezj, and M. Benaissa, "Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy," Talanta, vol. 211, p. 120740, 2020.

1.7.2.2 Conference paper

- H. Khadem, **H. Nemat**, J. Elliott, and M. Benaissa, "Multi-lag stacking for blood glucose level prediction," in 5th International Workshop on Knowledge Discovery in Healthcare Data, vol. 2675, 2020, pp. 146–150.

1.8 Thesis structure

Chapter 1 provides an overview of the research context, including background on diabetes care, along with the motivations, objectives, and contributions of the research presented in this thesis. A list of publications related to this thesis is also provided. In Chapter 2 a review of different aspects of developing BGL prediction models is provided. Then, the state-of-the-art AI-based techniques applied for BGL prediction are discussed. The contributions made in this thesis are presented in Chapters 3, 4, 5, and 6. These chapters are slightly modified versions of the published or submitted journal articles listed in Section 1.7.1.1.

Chapter 3 first describes the dataset used for developing different methodologies in this thesis. Then, it presents the methodologies for developing advanced prediction models leveraging deep learning and ensemble learning where novel meta-learning approaches are developed. The proposed models are then experimentally validated and discussed. In Chapter 4 the relations between BGL and affecting variables in T1DM management in the causality context are first investigated. Then, novel methodologies for leveraging causality information as prior knowledge for BGL prediction are developed. The methodologies are then evaluated using evaluation criteria and statistical analyses.

The research work presented in Chapter 5, develops and compares various prediction models integrated with PA. In this chapter different approaches for extracting PA information are proposed and the fusion of PA-related information with BGL data at different levels of fusion are explored. Comprehensive comparative analyses are then presented. In Chapter 6 the performance of BGL prediction using different data-driven forecasting approaches with univariate and multivariate inputs are compared. Regression-wised and clinical-wised metrics along with statistical analyses are then presented for evaluation and comparison purposes. Finally, Chapter 7 presents the thesis conclusions as well as some possible research directions.

Chapter 2

Blood glucose level prediction

2.1 Prediction model approaches

Based on the model structure and knowledge requirements, BGL prediction algorithms can be classified into three main groups: physiological models (extensive knowledge), data-driven models (black-box approach), and hybrid models (intermediate knowledge) [26, 42]. Among these approaches, the data-driven models, as the main focus of this thesis, gained popularity and various research has been increasingly performed to explore these approaches [43].

2.1.1 Physiological models

Physiological models aim to simulate the behaviour of real physiological systems using dynamic mathematical models. Hence, prior knowledge of physiological systems is a requirement for developing these models. These models are compartmental and derived by separating the body into uniform compartments. For BGL prediction, these models are designed to mathematically represent the dynamics of glucose-regulating systems. The glucose dynamics, the mechanism by which carbohydrate is converted to blood glucose, the process of insulin absorption, and the impacting model of PA on blood glucose regulation are the main compartments of a glucose-insulin physiological model for BGL prediction [5, 26]. These models are not particularly precise, and the physiological constants must be specified depending on prior information on glucose-relevant factors [44, 45].

2.1.2 Data-driven models

Data-driven models, which use experimental data and pattern recognition techniques to simulate glucose dynamics, have been proposed to overcome physiolog-

ical models' limitations [46, 47]. Data-driven, also known as empirical dynamic, models are black-box models generated from data only. These models, without any prior knowledge about the dynamics of glucose-regulating systems, can provide accurate predictions of glucose dynamics by determining the relation between the past, present, and future BGL. These models' advantages include no need for physiological information, minimal user interaction, and ease of development [43, 48]. Due to rapid advancements in data-driven AI methods, data-driven models have attracted considerable attention and are being increasingly explored [26, 45]. These models could be mainly classified into classical time series forecasting and machine learning approaches.

2.1.2.1 Classical time series forecasting

Classical time series forecasting approaches have also been used for the BGL prediction task [49]. Autoregression (AR) [50], autoregressive moving average (ARMA) [51], autoregressive integrated moving average (ARIMA) [52], autoregression with exogenous variables (ARX) [53], autoregressive moving average with exogenous variables (ARMAX) [54], and autoregressive integrated moving average with exogenous variables (ARIMAX) [55] are the common approaches used for BGL prediction. In the first three models, it is presumed that the future BGL would be a linear function of the historical BGL data, whereas the second three models incorporate exogenous variables into the univariate counterpart models.

2.1.2.2 Machine learning algorithms

The use of machine learning algorithms for time series forecasting has become increasingly popular in recent years. Time series forecasting can be restructured to supervised learning by converting data to a number of samples with input and output components. In this way, standard machine learning approaches can be used. Time series data can be transformed into samples with current and lag observations as input and future observations as output using a rolling window [20]. In the literature, various machine learning algorithms have been developed for BGL prediction. Most of these algorithms include ANNs [56, 57], decision trees [58, 59], kernel-based algorithms [29, 60], and regression techniques [61, 62]. According to reviews done by Mujahid et al. [18] and Woldaregay et al. [19], a majority of the data-driven approaches for BGL prediction in the literature used ANNs.

2.1.3 Hybrid models

Hybrid models combine both physiological and data-driven models to develop a BGL prediction model. Physiological models are often used as inputs for data-driven models, and the data-driven model component captures the association between the output of the physiological models and future BGL. Models of glucose dynamics [63], insulin dynamics [29], glucose-insulin dynamics [64], and meal absorption dynamics [65] are the most common physiological model components for developing hybrid models [5, 43].

2.2 Prediction model inputs

Data-driven models require accurate and large enough datasets. Data can be acquired from clinical trials or diabetes patient simulators. These models generally use information from historical BGL data, with or without other inputs.

2.2.1 Data origin

There are two types of data for developing BGL prediction models, real data from clinical trials and in silico data from diabetes patient simulators. Clinical datasets are the most commonly used type of data for BGL prediction [18], and more than half of them were collected in free-living situations [45].

The Ohio T1DM dataset [66, 67] with the capability of replication is the most widely used publicly available clinical dataset. Also, DirecNet [68], in which different protocols and data have been collected from children and adolescents with T1DM, is the second most frequently used clinical dataset in the literature [45].

Also, according to the review performed by Woldaregay et al. [19], AIDA [69] and UVa/Padova [70] were introduced as the two most used simulators for generating diabetes-related data. These simulators are usually preferred for evaluating the effectiveness of newly developed strategies for diabetes management prior to clinical research [18].

2.2.2 Input variables

Common inputs of BGL prediction models are the current and past information on BGL, carbohydrates, bolus, and PA. It is noteworthy that depending on the type of PA bands used, there are different kinds of PA data. These include heart rate, the magnitude of acceleration, step counts, galvanic skin response, skin temperature,

electrocardiogram, and electronic health record [18, 43]. Based on the review performed by Woldaregay et al. [19], the most commonly used group of variables used for BGL prediction is BGL, carbohydrate, and bolus. The second most frequently used set of variables includes BGL, carbohydrate, bolus, and PA. Using only BGL data ranked as the third most commonly used input.

BGL prediction from CGM data alone facilitates practical application in the real world, therefore there is no need to acquire and process data from multiple sensors and modalities. Hence, some work used BGL data only for developing data-driven prediction models [56, 71, 72, 73, 74, 75, 76]. While the BGL prediction model could be more reliable and accurate by considering other variables. Hence, others used BGL data along with other variables [57, 77, 78]. Investigating different inputs to find out if other variables can contribute to better prediction would be beneficial. Hence, some attempts have been made in this regard, however, a consensus has not been reached.

Zecchin et al. [40] showed that adding carbohydrate and bolus data to CGM data can improve the performance of BGL prediction using a neural network in a prediction horizon between 30 and 120 minutes. Also, Hameed et al. [41] concluded that whilst adding more information about carbohydrates and bolus adds more perturbations it does not always improve the accuracy of prediction. Jeon et al. [79] explored the impact of 19 physiological and monitoring variables provided in the Ohio T1DM dataset. By grouping the variables into four classes and creating 15 combinations of these groups, they concluded that using all feature classes could benefit BGL prediction by evading probably lost information.

2.3 Prediction horizons

The prediction horizon is the time the prediction model can provide BGL in advance. Prediction horizons can be classified into short-term (ranging from 15 to 60 minutes), medium-term (ranging from 90 to 240 minutes), and long-term (ranging from 360 minutes to one week) [18]. The widely used horizons range from 15 to 120 minutes, with acceptable accuracy for 15 and 30 minutes [19]. Moreover, in the two Ohio BGL prediction challenges, participants were asked to predict BGL 30 and 60 minutes in advance, hence these two horizons may be considered standard prediction horizons and were mostly used by researchers [39, 56, 79] in the literature.

2.4 Prediction performance assessments

BGL prediction models should be properly evaluated. For assessing BGL prediction performance, there are various evaluation criteria. Moreover, to have a conclusive validation for comparing two or more prediction models for several data contributors, statistical analyses should be conducted.

2.4.1 Evaluation criteria

The performance of developed prediction models can be evaluated using regression-wised criteria (also known as empirical accuracy) and clinical-wised criteria (also known as clinical accuracy). Regression-wised criteria evaluate the mathematical accuracy of prediction models without considering clinical significance, while clinical-wise criteria consider clinical usability significance.

2.4.1.1 Regression-wised evaluation

Primary metrics of regression-wised evaluation to calculate the overall performance of developed BGL prediction models include root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) which are calculated as Equations 2.1, 2.2, and 2.3, respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{2.1}$$

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.3}$$

In these equations, N represents the size of the testing set, and y_i, and $\hat{y}_i$ represent the reference and corresponding predicted BGL values, respectively. Also, in Equation 2.3, $\bar{y}$ denotes the mean of reference BGL values.

2.4.1.2 Clinical-wised evaluation

Error grid analyses and Matthews correlation coefficient (MCC) have been frequently utilised to have clinical insight regarding the performance of developed BGL prediction models. Error grid analysis including Clarke error grid (CEG) [80], Parkes error grid (PEG) [81], and surveillance error grid (SEG) [82], analyse and visualise BGL predictions using the comparison with the BGL reference values. SEG, which assigns risk values to each predicted BGL, is the most recently developed analysis. Also, a surveillance error (SE), which has been defined as the average of a bilinear interpolation of the SEG, is considered to have a unique clinical score for each patient [56].

MCC, which is a classification metric, has been deployed for clinical evaluation of BGL prediction in the literature [83, 84, 85]. The MCC criterion is used to assess the effectiveness of the models in distinguishing between adverse glycaemic (hypoglycaemia (BGL < 70mg/dL) or hyperglycaemia (BGL > 180mg/dL)) and normoglycaemic (70mg/dL < BGL < 180mg/dL) events. Accordingly, adverse glycaemic and normoglycaemic events are considered positive and negative classes, respectively. The predictions of the regression models are used to assign a prediction label. A confusion matrix is generated following comparing reference and predicted labels. The confusion matrix (Table 2.1) comprised true positives (TP), the number of adverse glycaemic events correctly predicted as adverse glycaemic events; true negatives (TN), normoglycaemic events correctly predicted as normoglycaemic events; false positives (FP), the number of normoglycaemic events incorrectly predicted as adverse glycaemic events; and false negatives (FN), the number of adverse glycaemic events incorrectly predicted as normoglycaemic events. MCC is then calculated as Equation 2.4.

Table 2.1: Confusion matrix for distinguishing between adverse and normoglycaemia events.

		Reference	
		Adverse	Normal
Prediction	Adverse	TP	FP
	Normal	FN	TN

Note. TP: True positive; FN: False negative; FP: False positive; TN: True negative.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.4}$$

2.4.2 Statistical analyses

Comparing the averages of evaluation criteria over different data sets is not meaningful [86]. Since there is a considerable variation between people with T1DM regarding their BGLs [19], for a more valid comparison between different BGL prediction models, statistical analyses need to be considered. There is limited work in the literature on conducting statistical analyses for prediction models [85, 87, 88].

2.4.2.1 Comparing two prediction models

To compare a newly developed prediction model with an existing or baseline model, a non-parametric Wilcoxon signed-ranks test [89] can be used, which is an appropriate test for comparing two approaches across multiple datasets with no assumption of normality.

2.4.2.2 Comparing more than two prediction models

To compare more than two prediction models, firstly, a Friedman test [90], which is the non-parametric counterpart of the ANOVA test [91], is conducted. If there is a significant difference between at least two prediction models, a post-hoc test such as Nemenyi test [92] is then performed to determine which models are performed significantly differently in a pair-wise fashion.

Also, to visualise the post-hoc results, a critical difference (CD) diagram [86] can be employed. A CD diagram is a graphical tool for comparing the outcomes of multiple models across multiple datasets. It displays the CD value which represents the minimum significant difference between pairs of models. In this diagram, various models are presented in order according to their average ranks and a horizontal line connects groups of not-significantly different models. Hence, it helps to identify which models are significantly distinct from each other and which are not.

2.5 Applications of advanced AI techniques

The literature shows an apparent acceleration in the utilisation of AI approaches for BGL prediction models [93, 94]. Improving the performance of BGL prediction is challenging, and even a small improvement is appreciated [95, 39, 96, 32, 97]. Recently, the majority of studies have tried to incorporate new advanced AI strategies to investigate their capabilities to enhance the performance of BGL prediction. In this section, some emerging AI techniques applied in BGL prediction, including deep learning, transfer learning, ensemble learning and causality analysis are

mentioned and related research in the literature is discussed. This section also summarises the results of different state-of-the-art methods for developing BGL prediction in the BGL prediction challenge. Figure 2.1 illustrates a schematic diagram of the factors involved in the development of a BGL prediction model.

2.5.1 Deep learning

Deep learning is a new area in AI which is inspired by the function of neurons inside human brains. In recent years, due to the growth of computing capability, deep learning models became more attractive. Some of the most important network types are feed-forward neural networks (FNNs) [98], convolutional neural networks (CNNs) [99], and recurrent neural networks (RNNs) [100].

FNNs are simple types of deep learning models which pass data in just one direction from the input layer to the output layer. These networks calculate the sum of the weighted inputs to find a mapping to output values. CNNs have convolutional layers. These types of layers consist of filters that are convolved with the input to extract local information from data [101],

RNNs are specially designed for time-dependent and sequence analyses. These networks have memory and feedback and the output of a layer can be fed back to the input [102]. RNNs suffer from the problem of vanishing gradients, which hampers the learning of long data sequences. The gradients carry information used in the RNN parameter update, and when the gradient becomes smaller and smaller, the parameter updates become insignificant, which means no real learning is done [102]. To overcome the vanishing gradient problem in long data sequence analysis, LSTM [103] and Gated recurrent units (GRUs) [104] were proposed. LSTM networks, instead of neurons, have memory blocks containing forget, input, and output gates that control the state of blocks [103]. GRUs use update and reset gates to control the output [104].

Deep learning models could be more effective at detecting complicated systems' dynamics rather than traditional machine learning approaches and have shown promising results. In BGL prediction, where the ability to capture the physiological dynamics of glycaemia is vital for accurate prediction, different deep neural network configurations have been successfully developed [15, 38, 105]. Also, deep learning models have evolved into more advanced BGL prediction paradigms and have earned a distinct place in the 2018 and 2020 Ohio BGL prediction challenges [93, 94]. The following provides an overview of some recent studies related to deep neural network models for BGL prediction.

In their study, Mirshekarian et al. [95] investigated several experiments for BGL

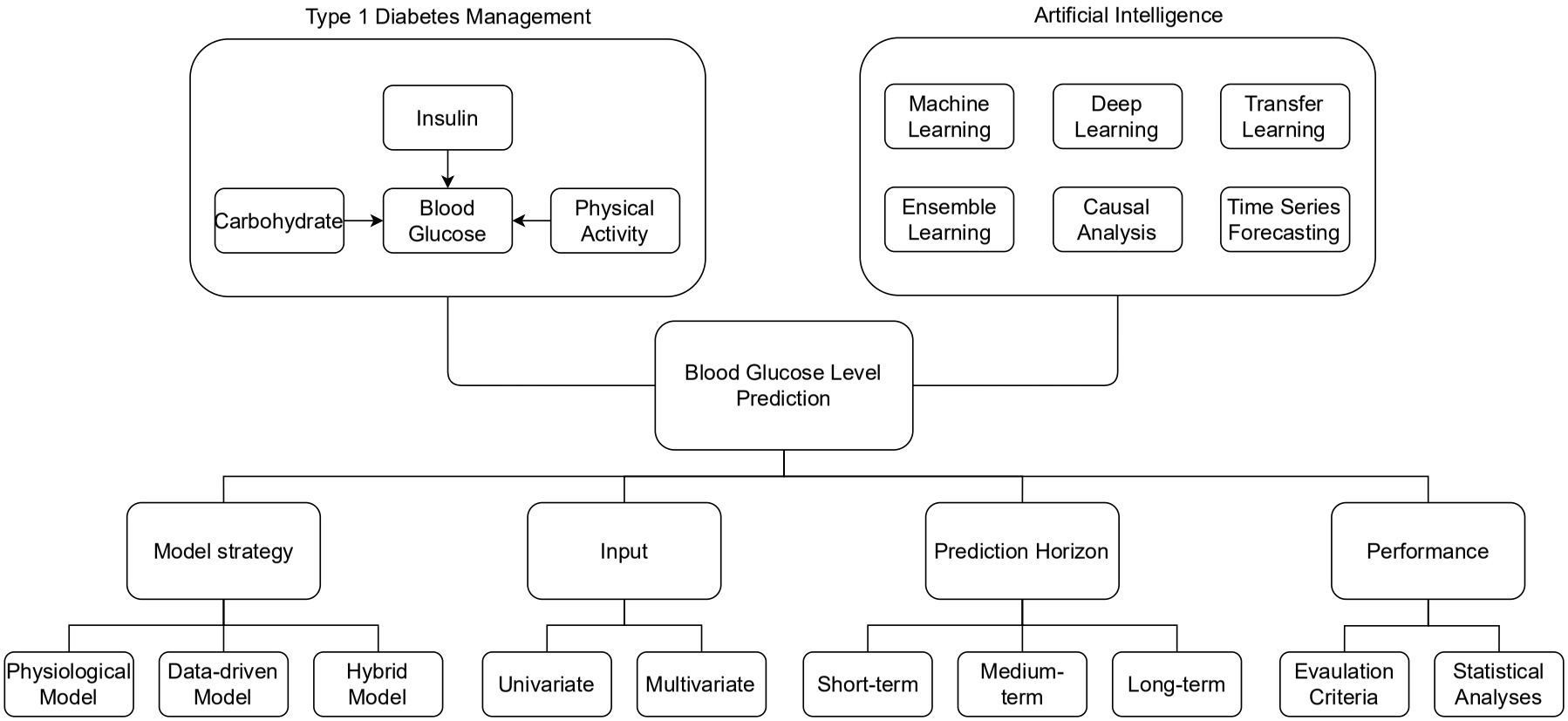


Figure 2.1: A schematic diagram for blood glucose level prediction.

prediction using CGM, insulin, meal, and activity data. They used the data of two diabetes simulators (i.e., AIDA and UVa/Padova) as synthetic datasets and the Ohio T1DM dataset as the real one and developed a new memory-augmented LSTM for the time series forecasting task in prediction horizons of up to one hour. They also considered an ARIMA model as a baseline and observed that the LSTM model meaningfully outperformed the baseline model. Based on the comparison results of the experiments for in-silico and real data, they found that the designed neural attention module improved prediction performance in synthetic data, although it failed to improve it in real data. Contrarily, using day time as an extra input of the LSTM model enhanced BGL prediction performance only in real data. They concluded that the behaviour of synthetic and real data is not always the same.

Li et al. [83] developed a deep learning model using a convolutional RNN architecture for BGL prediction. The model was comprised of convolutional and pooling layers, followed by a fully connected layer. They used BGL, bolus, and carbohydrate data from 10 virtual patients from the UVa/Padova simulator and 10 real people with T1DM as input. They evaluated and compared the proposed model with four baselines, including support vector regressions (SVR), an ARX, a latent variable model, and an ANN. The results showed the superior performance of the model compared to the baseline models.

Also, Martinsson et al. [56] developed a model using RNNs for predicting BGL with prediction horizons of 30 and 60 minutes using an end-to-end approach without the requirement of preprocessing or feature engineering. Their model was developed and evaluated using the Ohio T1DM dataset by considering the BGL data as input. A univariate Gaussian distribution was also used to estimate the certainty of the predictions. They evaluated their model using RMSE and SE metrics. Their method outperformed the naive baseline model.

Zhu et al. [106] developed a deep CNN as a modified version of WaveNet [107] for BGL prediction in people with T1DM. They categorised BGL prediction values into 256 classes; hence, the BGL prediction was converted from a regression task to a classification task. The classification model was mainly constructed by casual dilated CNN layers. They used BGL, insulin, carbohydrate, and the time index data from the Ohio T1DM dataset as input. Their results showed that their developed model differed from existing RNN models and compared to many current algorithms performed better. Also, they later proposed a model using dilated RNNs for predicting BGL 30 minutes in advance. They investigated vanilla RNNs, LSTMs, and GRU architectures before selecting a vanilla RNN for the final model. Their model was trained using BGL, bolus, and meal intake data from the Ohio T1DM

dataset and the UVa/Padova simulator. Compared to the Ohio dataset, their proposed model performed better for BGL prediction in the synthetic dataset. Furthermore, their findings indicated that preprocessing steps such as interpolation and extrapolation could improve the performance of prediction. Their model had a smaller RMSE compared to AR, SVR, and CNNs. Hence, they expressed that the dilated RNN model could improve the performance of BGL prediction [77].

2.5.2 Transfer learning

In artificial intelligence, transfer learning refers to the use of prior experiences to improve learning. Transfer learning involves fine-tuning a pre-trained model to accomplish a related task more effectively. This field has been successfully deployed in different areas, including computer vision, natural language processing, and healthcare [22, 96, 108, 109]. Several studies have examined the efficacy of transfer learning in BGL prediction in recent years. Some of the recent work deploying transfer learning in BGL prediction is briefly described below.

Bhimireddy et al. [110] developed several sequence-to-sequence multivariate ANN architectures, including LSTM, BiLSTM, convolutional LSTMs, temporal convolutional networks, and sequence-to-sequence models for BGL prediction in T1DM. A gradient boosting algorithm was also used for selecting important features of the data in order to develop transfer learning models. They developed and evaluated their models using the Ohio T1DM dataset. Their results showed that sequence-to-sequence models outperformed transfer learning. Also, Zhu et al. [77] applied transfer learning to exploit other subjects'data for training each individual model. They found it useful for one subject with various missing data.

Daniels et al. [96] investigated the effectiveness of multitask learning as a type of transfer learning, in BGL prediction. They developed single-task learning, transfer learning, and multitask learning using a convolutional recurrent neural network. These approaches were compared with an SVR model, as a baseline model, and also with each other. They also considered BGL variability, as proper knowledge for dividing the experiment into two groups to perform multitask learning. They used the OhioT1DM dataset for developing and validating their models. Results showed that the developed multitask learning approach outperformed other models for short-term and long-term prediction horizons. They concluded that multitask learning can be deployed for personalised models on limited individual data to promote BGL prediction.

De Bois et al. [111] developed a multi-source adversarial transfer learning architecture for enhancing data transfer quality between different sources. Their archi-

tecture allowed for the learning of a feature representation consistent across sources, making the learning process more universal and transferable. They utilised an SVR and two fully CNNs as baseline models. They also compared the proposed adversarial transfer models with standard transfer models. They used three different sources of data including T1DM patients, T2DM patients, and a T1DM simulator. Their developed multi-source transfer learning could help with the lack of big-enough data for training deep learning and improve the performance of BGL prediction.

Shuvo and Islam [97] incorporated multitask learning into a deep learning model to predict BGL in T1DM. Their proposed architecture was comprised of two layers of stacked LSTM, as shared hidden layers, and two dense layers, as clustered hidden layers. These were followed by subject-specific dense layers. For developing their model, they used the Ohio T1DM dataset and evaluated it using RMSE, MAE, and CEG. The results showed an enhancement caused by multitask learning compared to other machine learning and deep learning models.

2.5.3 Ensemble learning

Ensemble methods, as one of the advanced AI approaches, learn from several machine learning models, inferred as base-learners. The core assumption of ensemble learning is that improvements could happen due to the compensation of the single base-learner's error by other base-learners. Ensemble models are constructed in two main steps, generating base-learners and integrating base-learners. Considering base-learners' generation, ensemble methods can be categorised as homogeneous and heterogeneous. In homogeneous ensembles, the base-learners are generated by a single algorithm, whereas in heterogeneous ensembles, at least two distinct algorithms are used. Base-learners' combination, also inferred as output fusion, is the process of combining outputs from base-learners. The two main approaches for output fusion include weighting methods and meta-learning methods. Base-learners' outputs can be weighted and averaged to make a single output [31, 32]. In meta-learning algorithms, there are two levels of learning. At the first level, multiple base-learner models are trained, and the predictions of the base-learner models are combined to make the final prediction at the second level. Bagging [112], boosting [113], and stacking [114] are three known meta-learning algorithms for constructing ensemble models.

Since glucose dynamics are complex, ensemble techniques can be used to improve the accuracy of BGL prediction by combining multiple models [19, 32]. Recently, several researchers have examined the use of different ensemble models for predicting BGL. Some of the recent work is briefly described in the following.

Jeon et al. [79] performed an investigation for predicting BGL in the prediction horizon of 30 minutes using the Ohio T1DM dataset. Following their previous work [115], in which they showed that their developed gradient-boosted regression tree (XGBoost) model, consisting of an ensemble of decision trees with a gradient boosting framework, outperformed a random forest regression and an LSTM model, they developed five variants of the model using optimised parameters. They further combined the predictions from five models using weighting output fusion models to generate an ensemble model and demonstrated that the ensemble model made better BGL predictions compared to the individual models.

Saiti et al. [116] developed three ensemble models using ARX and SVR as base-learners, followed by linear, bagging, and boosting meta-learning output fusions. BGL, carbohydrate, and insulin were used for developing prediction models. The models were evaluated and compared for prediction horizons of 30, 45, and 60 minutes. Their results showed that the three ensemble models outperformed both individual learners.

In the Ohio BGL prediction challenge 2020, we developed two methods leveraging ensemble learning for BGL prediction in T1DM using BGL and PA data. Three heterogeneous base-learners including an multilayer perceptron (MLP), an LSTM, and a partial least squares regression (PLSR) were used. In one method histories of BGL data appended with the average of PA in the same histories were used to train base-learners. In the other method, histories of BGL and PA were used separately to train the same base-regressions. The predictions from the base-learners were used as input for a PLSR model to create a combined model using a stacked meta-learner to make the final predictions. The results showed the effectiveness of both methods for BGL prediction [117]. This initiated proposing novel meta-learning approaches [71] which are presented in Chapter 3.

Also, in the same challenge, using the Ohio T1DM dataset, Khadem et al. [118] developed ensemble models for BGL prediction utilising different lags. Three heterogeneous linear and ANN base-learners were trained using both 30 and 60 minutes of historical BGL data. These models were then used to create two uni-lag and one multi-lag system. In the uni-lag system, base-learners were trained using either 30 or 60 minutes of historical BGL data, whereas, in the multi-lag system, base-learners were trained using both lags. In all three systems, a PLSR model was used as a meta-learner based on stacked meta-learning. The results showed that the stacking systems outperformed the individual models. Also, the multi-lag system achieved the best prediction accuracy.

Also, Wadghiri et al. [32], in their review verified that homogeneous ensembles,

mostly followed by bagging and boosting meta-learning, improved the performance of BGL prediction compared to individual machine learning models, still comparable to deep learning models. While heterogeneous ensembles, mostly followed by stacking meta-learning, outperformed single machine learning and deep learning models. Hence, heterogeneous ensembles could be generally considered better than homogeneous ones in BGL prediction. Also, ANNs were determined as the most widely used base-learners for constructing heterogeneous ensembles, according to their review.

Furthermore, Khadem et al. [85] proposed a lag fusion framework employing meta-learning analysis to address the challenge of determining an appropriate history length for model training in the BGL prediction task. The developed method utilised MLP and LSTM models trained using four history lengths of 30, 60, 90, and 120 minutes. Then, an interconnected lag fusion approach was developed based on nested ensemble learning. The analyses were performed using the Ohio T1DM dataset. The results showed that the proposed method was effective in BGL prediction.

Langarica et al. [119] proposed an ensemble model using meta-learning output fusion for personalised BGL prediction. They used synthetic data from the UVa/Padova simulator and showed that their model needed fewer data and training iterations compared to a transfer learning approach. Their model outperformed baselines, especially for longer prediction horizons.

2.5.4 Causal analysis

Causality analysis studies the cause and effect relations. One main approach for causality analysis is casual inference where causal relations are quantified [120, 121]. In recent years, data-driven causal inference approaches have emerged to determine causal relations between variables in time series data [122]. In these approaches, causation related to observations of variables in a complex system is assessed as causality from multivariate time series [123, 124].

Causality investigation approaches from time series data can be categorised into four main approaches [125, 126]: regression-based approaches, such as Granger causality [127], which use histories for predictions; information theoretic approaches, such as transfer entropy [128], which use conditional mutual information; state space dynamics based approaches, such as convergent cross mapping [129]; and graphical approaches, such as Peter Clark momentary conditional independence algorithm [130, 131], which estimate causal inference in high-dimensional time series.

Also, causality has recently found applications in time series forecasting tasks in different fields including neuroscience [33], climate [35], and economic data [132, 133]. These applications tried to improve the performance of time series forecasting by selecting affecting features using causality.

There is a lack of causality analysis deployment in BGL prediction in the literature, and developed prediction models with causal neural networks have been the only way to deploy causality in BGL prediction [106, 134]. To contribute to fill this gap, we used causal information as prior knowledge and developed causality-informed prediction models [87] which is presented in Chapter 4.

2.5.5 The prediction challenge

The second BGL prediction challenge was held at the fifth international workshop on knowledge discovery in healthcare data in August 2020 [135]. The participants were asked to predict BGL for the prediction horizons of 30 and 60 minutes using the newly released Ohio dataset [67]. The evaluation results had to be reported as RMSE and MAE for both prediction horizons. The performance of different prediction approaches on the Ohio_2020 dataset [67] was compared and ranked. Table 2.2 summarises the methods and the evaluation results of the top 10 papers for BGL prediction.

Table 2.2: Comparison of the evaluation results of blood glucose level prediction challenge

Paper	Method	PH: 30 min		PH: 60 min	
		RMSE	MAE	RMSE	MAE
Freiburghaus et al. [136]	CRNN	17.45	11.22	33.67	23.25
Rubin-Falcone et al. [137]	DRNN	18.22	12.83	31.66	23.60
Hameed and Kleinberg [138]	VRNN	19.21	13.08	31.77	23.09
Zhu et. al. et al. [139]	GAN	18.34	13.37	32.21	24.20
Yang et al. [140]	MS-LSTM	19.05	13.50	32.03	23.83
Bevan and Coenen [141]	LSTM	18.23	14.37	31.10	25.75
Sun et al. [142]	LV	19.37	13.76	32.59	24.64
Khadem et al. [118]	MLS	19.01	13.73	33.37	24.98
Nemat et al. [117]	Stacking	18.99	13.73	33.39	25.04
Daniels et al. [143]	MTCRNN	19.79	13.62	33.73	24.54

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; CRNN: convolutional recurrent neural network; DRNN: Deep residual neural network; vanilla recurrent neural network; GAN: Generative adversarial network; MS-LSTM: Multi-scale long short-term memory network; LSTM: Long short-term memory; LV: Latent variable, MLS: Multi-lag stacking; MTCRNN: Multitask approach using convolutional recurrent neural network.

2.6 Benchmark of prediction models

It is also of interest in the literature to compare the efficacy of different prediction models as a factor in the advancement of BGL prediction. The use of different datasets, inputs, and model settings has made this comparison difficult to be conclusive, and limited number of studies have been conducted on this topic. Xie and Wang [39] benchmarked a classical ARX model against 10 different machine learning approaches for predicting BGL in T1DM. These models included Elastic-Net, Lasso, Huber, Random-Forest, Gradient-Boosting-Trees, Ridge, and SVR, with both linear and radial basis kernels, along with two deep learning models (i.e., vanilla LSTM and temporal convolution networks). They used the Ohio dataset and considered BGL, insulin, carbohydrate, and exercise data as input. Their results showed that the ARX model and Ridge regression had the lowest average RMSE in the prediction horizon of 30 minutes for BGL prediction. However, the ARX model had worse robustness compared to the DNNs. It over-predicted peaks while under-predicting valleys.

Rodriguez et al. [144] to contribute to better management of T1DM, developed and compared four machine learning approaches, including a Gaussian process with radial basis function kernels, an MLP, an SVM, and a Bayesian regularised neural network. They used glycaemia-related data, including BGL, insulin, meal, step count, heart rate, and sleep. The data was collected under real-world conditions from 25 people with T1DM over a 14-day monitoring period within the context of the Internet of Things. They showed that the Bayesian neural network performed best on R^2 and RMSE metrics and introduced it as the most capable technique for modelling BGL dynamics.

Moreover, Zhang et al. [38], compared four different data-driven models to forecast BGL in T1DM. The models included a dilated CNN model, a sequence-to-sequence LSTM model, a bidirectional reservoir computing model, and a newly developed multiple linear regression model. They used multiple variables measured by sensors or self-reported in the Ohio datasets. They found that while their sequence-to-sequence LSTM model was the most accurate at predicting BGL 30 minutes, in advance, their multiple linear regression model performed best at predicting BGL 60 minutes, in advance, at a lower computational cost.

2.7 Recent research in physical activity

PA is a determinant of insulin sensitivity and an important factor in T1DM management [10, 36, 37]. Several investigations have been performed examining PA in T1DM management. As part of our review, we categorise the relevant PA-involved works in T1DM into three main areas: detection, classification, or description of PA [145, 146, 147, 148], investigation of the impact of PA on T1DM management [30, 149, 150], and inclusion of PA in adverse glycaemic events detection and BGL prediction [37, 151, 152].

In relation to the analysis of PA itself, in a study by Cho et al. [145], accelerometer, heart rate, and CGM data were used to detect and classify the type and intensity of PA by developing random forest models. Also, Cescon et al. [146] detected and classified PA based on its intensity by developing deep learning models using accelerometer data collected from wristbands in free-living conditions. Moreover, Dénes-Fazakas et al. [147] by investigating different machine learning models, detected the presence of physical activity from CGM and HR data. Also, Ozaslan et al. [148] proposed a physiological model for activity on board using step counts.

Machine learning approaches have been used in some studies to investigate the impact of PA on managing T1DM. By analysing data from 37 T1DM patients using linear regression models, Ozaslan et al. [149] concluded that PA can have immediate and delayed effects on BGL. Also, they found that there is a significant relationship between PA and BGL after an evening meal, suggesting that measuring PA may be helpful for guiding meals. Also, Ozaslan et al. [150] proposed an insulin dosing system by adding PA information, which significantly decreased time spent in hypoglycaemia and increased time spent in normoglycaemia. In their study, Tyler et al. [30] collected a dataset consisting of highly-controlled exercise sessions and investigated several machine learning models to quantify the effect of physical activity on BGL. They developed an adaptive personalized machine learning model to predict BGL changes related to exercise.

Some studies performed to predict BGL or glycaemic events by including PA. Xie and Wang [37] developed a glucose dynamics model by considering PA. They proposed a non-linear autoregressive moving average model with exogenous inputs by entering the PA. To train and evaluate the model, they used in silico data from UVa/Padova simulator. They observed that during and two hours after exercise, the nonlinear and linear models with PA made a better prediction for BGL in a prediction horizon of 30 minutes, rather than the linear model without PA. Also, Bertachi et al. [152] investigated the possibility of nocturnal hypoglycaemia prediction in T1DM by incorporating PA. They analysed the data of CGM sensors and physical

activity trackers collected in 12 weeks from 10 people with T1DM. They applied MLP and SVM models for binary classification. They concluded it was feasible to predict nocturnal hypoglycaemia from CGM and activity data using machine learning approaches. Moreover, Hobbs et al. [151] developed a glycaemic model by considering some terms indicating different effects of PA on metabolism. They showed their model outperformed the prediction model using only BGL.

2.8 The Ohio dataset

For developing and evaluating our methodologies, we used the publicly available Ohio T1DM dataset [66, 67]. The Ohio T1DM dataset comprises two sets of data from 12 people with T1DM, in total. The first dataset related to six T1DM patients was released in 2018 for the first BGL prediction challenge [153] (called Ohio_2018), and the second dataset related to an additional six patients was released in 2020 for the second BGL prediction challenge [135] (called Ohio_2020).

Each dataset contains eight weeks of data. There are two separate XML files for each participant for training and testing subsets, with the last 10 days' worth of data as testing, and the rest as training. The datasets provide data from physiological sensors and self-reported life events. All patients were on Medtronic 530G or 630G insulin pump therapy with a Medtronic Enlite CGM sensor to monitor and collect their BGL data every five minutes. Participants wore the Basis Peak fitness band data with 5-minute aggregation and Empatica Embrace bands with 1-minute aggregation, for the dataset released in 2018 and 2020, respectively. Also, data contributors reported carbohydrate estimations and meal times along with other life events including sleep, work, and illness. Table 2.3 presents the information related to gender, age, and the number of data points for each variable, for both datasets.

Table 2.3: Gender, age, and the number of data points in training and testing sets related to the contributors in the Ohio_2018 and Ohio_2020 datasets.

	PID	Gender	Age	Training data points				Testing data points			
				BGL	Carb	Bolus	HR	BGL	Carb	Bolus	HR
Ohio_2018	559	female	40–60	10796	150	152	11979	2514	29	36	2633
	563	male	40–60	12124	129	347	11966	2570	27	89	2706
	570	male	40–60	10982	136	326	12328	2745	33	84	2720
	575	female	40–60	11866	243	187	12446	2590	45	36	2698
	588	female	40–60	12640	221	182	12980	2791	37	40	2620
	591	female	40–60	10847	212	261	12276	2760	41	51	2668
	PID	Gender	Age	Training data points				Testing data points			
				BGL	Carb	Bolus	MA	BGL	Carb	Bolus	MA
Ohio_2020	540	male	20–40	11947	73	309	28353	2896	27	87	7682
	544	male	40–60	10623	159	134	53219	2716	38	39	13796
	552	male	20–40	9080	78	336	25604	2364	21	102	10382
	567	female	20–40	10858	32	313	39037	2389	0	54	9163
	584	male	40–60	12150	95	268	47024	2665	23	54	12737
	596	male	60–80	10877	265	208	34790	2743	54	38	8643

Note. PID: Patient identity; BGL: Blood glucose level; Carb: Carbohydrate intake; Bolus: Injected bolus insulin; HR: Heart rate; MA: Magnitude of acceleration.

Chapter 3

Leveraging ensemble learning in blood glucose level prediction

3.1 Preface

Ensemble learning is an advanced strategy that can enhance the performance of ML tasks by combining multiple models. The key idea of ensemble learning is that improvements can occur as a result of multiple base-learners compensating for the inaccuracy of a single base-learner. Using deep and ensemble learning together has emerged in recent years as an attractive strategy due to the growth of computing capability. Recently a number of studies have been performed combining deep learning models and the ensemble learning concept in the BGL prediction field [117, 118, 154]. However, there is still a lack of a comprehensive investigation of deep and ensemble learning capability and comparison with benchmark models.

The work presented in this chapter proposes new advanced architectures for BGL prediction in T1DM leveraging the combination of deep and ensemble learning. Two types of LSTM networks along with a linear regression model, are considered benchmark BGL prediction models. These benchmark models are also used as base-learners in the ensemble architectures. Three meta-learning approaches are developed for the output fusion of base-learners in advanced architectures. The performance of the proposed ensemble models is then compared with benchmark non-ensemble models.

3.2 Material and methods

3.2.1 Dataset

In the work presented in this chapter, we used BGL data from both Ohio T1DM datasets [66, 67] described in Section 2.8.

3.2.2 Preprocessing

The first step in the preprocessing was to deal with the missing data. Missing data in the training set were imputed using linear interpolation. Also, for the testing set, linear extrapolation was used in order to ensure that future data were not observed by the model and that the model can be used for a real-time application. So, BGL data were converted to a regular time series in 5-minute intervals without any missing data. For example, Figure 3.1 shows the first 1000 points of original and interpolated training data after data imputation for patient 575.

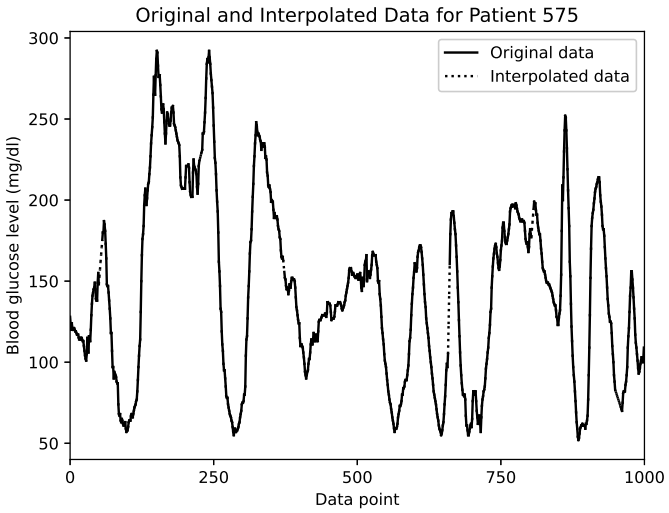


Figure 3.1: The first 1000 blood glucose level data points of the training set for patient 575 after interpolation.

Another data preprocessing step was to reframe the time series problem to a supervised learning task. In the current work, the task of BGL prediction was approached as a sequence-to-sequence problem, where we looked for predicting the future BGL sequence based on the historical sequence of BGL. To do so, time series data were transformed into samples with lag observations as input and future observations as output. Then a rolling window with four different history lengths of 6, 12, 18, and 24 data points was investigated for the input, which carried the information

of 30, 60, 90, and 120 minutes of history, respectively. The associated output was a vector with 6 and 12 data points corresponding to the 30- and 60-minute prediction horizons, respectively.

In the final step of preprocessing, input sequences were scaled to the minimum and maximum value over the entire training set of all subjects.

3.2.3 Prediction models

Linear models could be appropriate tools for BGL prediction tasks as they are simple and only require low-cost computing. On the other hand, LSTM networks, as a type of RNNs, which are suitable for working with sequential data and time series forecasting [20], are effective in predicting BGL [83, 95]. A linear regressor and two different types of LSTM networks were developed in the present work, followed by proposing three different approaches using ensemble learning. The following sections present a naive baseline model, three non-ensemble models, and finally, three ensemble models developed for the BGL prediction task.

3.2.3.1 Baseline model

A baseline model requiring a comparison level of performance is crucial for any time series forecasting task. Being simple, fast, and repeatable are three characteristics of a good baseline model [20]. In this work, a naive baseline model, considering the last available BGL value as the forecast, was used.

3.2.3.2 Non-ensemble models

One linear model and two types of LSTM networks, postulated as effective approaches in BGL prediction [83, 95, 155], were developed as prediction models.

Linear regression It is a simple and an easy-to-apply model with minimal computational cost. A linear regression model fits a model on the training dataset by minimising the error between real targets and predictions from a linear approximation [156]. Further, a simple linear model was developed for the BGL prediction task, and a linear model was fitted for each data contributor using the input and output vectors of the training set.

Vanilla long short-term memory (VLSTM) A vanilla LSTM network [103] with the vector output was used for multi-step ahead forecasting. The model was composed of an LSTM layer with 200 units, followed by a Dense layer with 100

units and an output layer with the number of future data points as the number of units (Figure 3.2). To train the model, the MSE was used as the loss function. The epoch size and batch size were considered as 500 and 32, respectively. The callback of ReduceLROnPlateau was employed for reducing the learning rate with the initial learning rate of 0.01 by a factor of 0.1 when validation loss has stopped improving for a patient number of 20 epochs. The initialiser, activation function, and optimiser were tuned for each history and horizon, which are discussed in the next section.

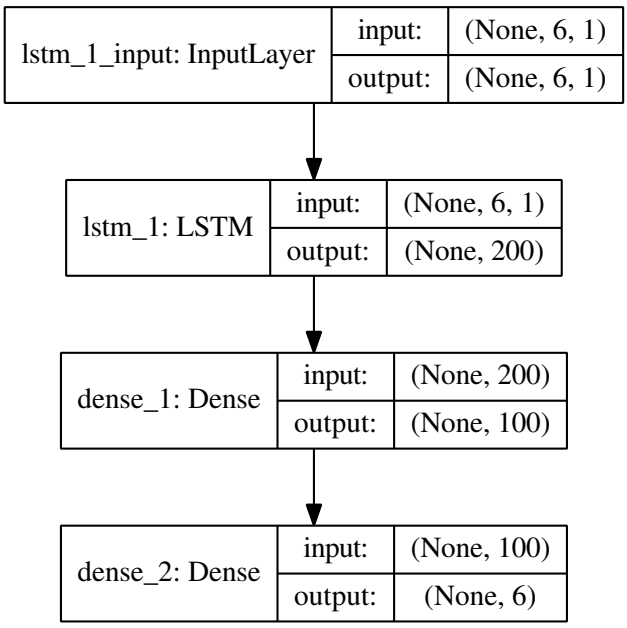


Figure 3.2: Plot of the VLSTM model.

Bidirectional long short-term memory (BiLSTM) BiLSTM is another type of RNNs that can be used for sequence forecasting tasks [20]. A BiLSTM model was implemented using a Bidirectional LSTM layer with 200 units, followed by a Dense layer with 100 units and an output layer. Similar to the VLSTM model, the loss function, epoch size, and batch size were considered as MSE, 500, and 32 for training the model, respectively. Moreover, ReduceLROnPlateau was employed as the callback with an initial learning rate of 0.01. The initialiser, activation function, and optimiser were optimised for each history and horizon, which is discussed in the following section.

Developing NN architectures for a given problem is challenging as no specific rules exist. Two categories of parameters need to be determined when developing a network: parameters related to the network's structure, including the number of hidden layers, the number of units in each layer, the initialiser, and the activation

function, and parameters related to compilation, including batch size, learning rate, and the optimiser.

Ideally, a network architecture should include as few hidden layers and units as necessary. Hence, we started with one hidden layer. Also, grid search and random search are two common approaches to determining some parameters. For determining initialiser, activation, and optimiser, due to computational limitations, we used random search.The Search Space for each parameter was selected based on the developed BGL prediction models in the literature.

To optimise hyperparameters, the training set was divided into training and validation subsets. For this purpose, the first 80% of data was allocated to the training set, and the following 20% was considered for the validation set. Then, the parameters were fine-tuned by selecting the ones that resulted in the lowest average RMSE over the validation data of 12 subjects. In addition, the hyperparameters were separately optimised for the prediction horizons of 30 and 60 minutes.

The length of the history window was the first parameter to be optimised. To do so, four history window lengths of 30, 60, 90, and 120 minutes which were commonly used values in the literature [56, 77] for tuning, were investigated. These history lengths included 6, 12, 18, and 24 history points, respectively. The two LSTM models were individually fine-tuned for each history length to have a fair comparison between all histories.

To tune the LSTM models, due to computational costs, the epoch size was amended to 200. The initialiser and activation function related to layer configuration and optimiser related to the compilation process were tuned. To tune each parameter, the two other parameters were fixed and the variable was changed over its search space.

To tune the VLSTM model for the prediction horizon of 30 minutes, the kernel initialiser was selected among {Glorot uniform and He uniform} by considering ReLU and Adam as the activation function and the optimiser, respectively. As a result, He uniform and Glorot uniform were selected for the history window of 30 and 60, as well as 90 and 120 minutes, respectively. Then, the search space of {ReLU and Tanh} was explored to tune the activation function by considering the selected initialisers for each history and Adam as the optimiser. It should be noted that ReLU was selected for all histories. The optimiser was the last parameter to be tuned while considering the selected values for the initialiser and activation function. Additionally, for all histories from the search space of {Adam and Adagrad}, Adam optimiser was chosen.

A similar process to that of the VLSTM was repeated for tuning the BiLSTM

model. For the prediction horizon of 30 minutes, Glorot uniform was selected for the history windows of 30 and 90 minutes regarding the prediction horizon of 30 minutes, and He uniform was chosen for the history windows of 60 and 120 minutes as the kernel initialiser. For all histories, ReLU and Adam were selected as the activation function and optimiser, respectively.

Similarly, the Glorot uniform was selected for the history windows of 30 and 120 minutes concerning the prediction horizon of 60 minutes in the VLSTM model, followed by choosing He uniform for the history windows of 60 and 90 minutes as the kernel initialiser. Further, ReLU and Adam were selected as the activation function and optimiser, respectively, for all histories. Regarding the BiLSTM model, Glorot uniform was chosen for the history windows of 30 and 120 minutes, and He uniform for history windows of 60 and 90 minutes as the kernel initialiser. Furthermore, ReLU and Adam were selected as the activation function and optimiser for all histories, respectively.

Eventually, using the validation set, the average RMSE over all patients for each history window was calculated and used as a criterion for choosing the history length. Figure 3.3a illustrates the results of this investigation for the prediction horizon of 30 minutes, and Figure 3.3b shows those for the prediction horizon of 60 minutes. The final chosen hyperparameters for VLSTM and BiLSTM models regarding both prediction horizons of 30 and 60 minutes are presented in Table 3.1.

According to Figure 3.3, two graphs related to both prediction horizons of 30 and 60 minutes were also compared for each model. As shown, the Linear graphs for both prediction horizons using the four different history lengths resulted in the same average RMSE, implying that the performance of this model is similar for these history lengths. It can also be interpreted as robustness for the Linear model. Considering the VLSTM graphs, due to various RMSE among different history windows, the history length could noticeably affect the performance of this model. For both prediction horizons, the history of 90 minutes led to the least averaged RMSE for this model thus, it was chosen for the history length regarding training the model. Considering the BiLSTM graphs, moderate variation could be observed among the four different history window lengths as well. The history length of 60 minutes was the best one for this model in both prediction horizons.

3.2.3.3 Ensemble models

Ensemble methods are advanced approaches for solving a range of machine learning tasks. In meta-learning data fusion approach, there are two levels of learning. At the first level, multiple base-learner models are trained, followed by combining the

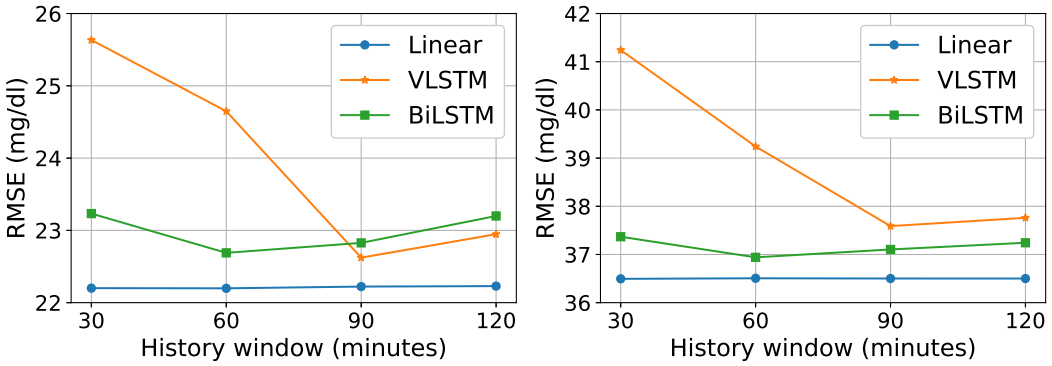


Figure 3.3: Tuning the length of the history window for the prediction horizon of 30 (a) and 60 minutes (b).
Note. RMSE: Root mean square error; VLSTM: Vanilla long short-term memory; BiLSTM: Bidirectional long short-term memory.

Table 3.1: Selected hyperparameters of the VLSTM and BiLSTM models.

Parameter	VLSTM		BiLSTM	
	PH: 30 min	PH: 60 min	PH: 30 min	PH: 60 min
History	90 minutes	90 minutes	60 minutes	60 minutes
Initialiser	Glorot uniform	He uniform	Glorot uniform	He uniform
Activation	ReLU	ReLU	ReLU	ReLU
Optimiser	Adam	Adam	Adam	Adam
Cell type	Vanilla LSTM	Vanilla LSTM	Bidirectional LSTM	Bidirectional LSTM

Note. PH: Prediction horizon; VLSTM: Vanilla long short-term memory; BiLSTM: Bidirectional long short-term memory.

predictions of base-learner models for making the final prediction at the second level [31].

This work looks into the second level of learning in three ways—i.e., univariate, multivariate, and two-dimensional data analysis. In the proposed methodologies, meta-learning output fusion was used at the second learning level to integrate base-learners outputs into one final prediction.

The non-ensemble models (i.e., Linear, VLSTM, and BiLSTM) were used as base-learners. The outputs of base-learners were used as the input of a meta-learner. To fuse the outputs of base-learners, stacking [114] and two novel approaches, named Multivariate and Subsequences, were investigated. The meta-learners were chosen for each approach based on the requirements of the output fusion in the second level of learning.

During reframing time series to sequence-to-sequence samples, non-equal history length of base-learners (30, 60 and 90 minutes for the Linear, BiLSTM, and

VLSTM models, respectively) resulted in generating twelve and six more samples for Linear and BiLSTM than VLSTM. The first twelve and six data points were discarded from training and testing subsets used for the BiLSTM and Linear models to equalise the sample sizes, which was an integration provision.

Stacking approach In this model, the output sequences of three base-learners were stacked and used as the input sequence of a meta-learner. VLSTM, BiLSTM, and Linear models were considered as base-learners, and by virtue of the simplicity the Linear model was regarded as the meta-learner. Figure 3.4a depicts the schematic of this approach for the BGL prediction of 30 minutes in advance where $\hat{Y_1}$ is the output sequence of the Linear model consisting of six points ahead prediction values of $\hat{y}_{11}$, $\hat{y}_{12}$, $\hat{y}_{13}$, $\hat{y}_{14}$, $\hat{y}_{15}$, and $\hat{y}_{16}$. Similarly, $\hat{Y_2} = [\hat{y}_{21},\ \hat{y}_{22},\ \hat{y}_{23},\ \hat{y}_{24},\ \hat{y}_{25},\ \hat{y}_{26}]$ and $\hat{Y_3} = [\hat{y}_{31},\ \hat{y}_{32},\ \hat{y}_{33},\ \hat{y}_{34},\ \hat{y}_{35},\ \hat{y}_{36}]$ represent the output sequences of VLSTM and BiLSTM models, respectively. These three output sequences were concatenated to feed the meta-learner. The output of the meta-learner was the final prediction.

Multivariate approach In this method, the outputs of the base-learners were considered as different variables. The existing univariate time series forecasting task at the first level of learning was converted to a three-variate time series forecasting task at the second level of learning. Considering the technique of meta-learning output fusion, a multivariate LSTM model was used as the meta-learner. Figure 3.4b illustrates a diagram of this methodology for the 30-minute prediction horizon. As shown, $\hat{Y_1}$, $\hat{Y_2}$, and $\hat{Y_3}$ (the output sequences of base-learners) were simultaneously used as a three-variable input sequence for the meta-learning process.

Due to similarities in the architectures of this model and the univariate VLSTM model and for a reduction in computational costs, the same hyperparameters tuned for the univariate model were used instead of performing a separate hyperparameter-tuning process. Hence, the model composed of an LSTM layer with 200 units followed by a fully-connected Dense layer with 100 nodes and an output layer. Both hidden layers used ReLU as the activation function. Glorot uniform and He uniform were used as the initialiser for the prediction horizons of 30 and 60 minutes, respectively. Furthermore, MSE and Adam were used as the loss function and the optimiser. The model was trained with 500 epochs with a learning rate of 0.01 and an epoch size of 32.

Subsequences approach In this method, the VLSTM, BiLSTM, and Linear models were used as base-learners. In this regard, we looked into their output sequences

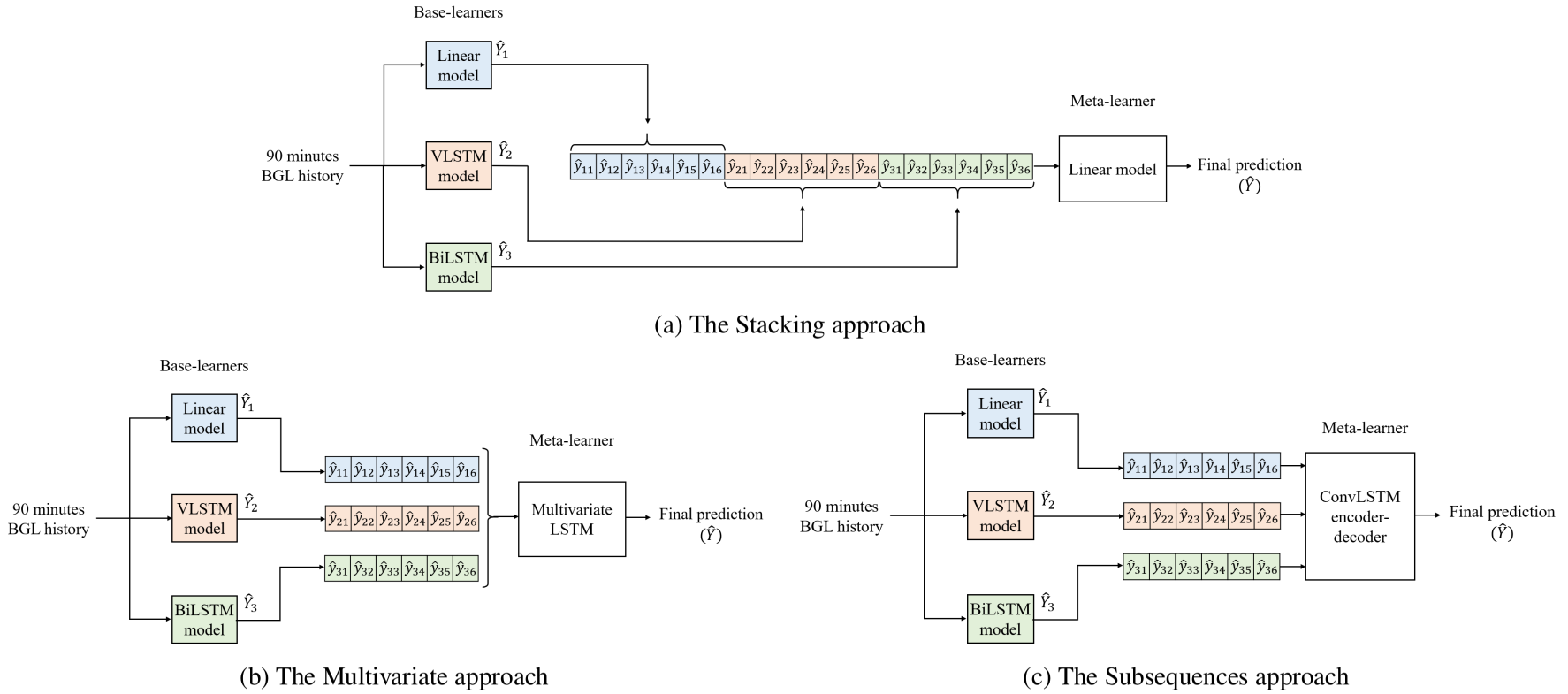
and considered $\hat{Y_1}$, $\hat{Y_2}$, and $\hat{Y_3}$ output sequences as three subsequences for the meta-learner. In this way, our one-dimensional time series forecasting task was configured as a two-dimensional data analysis problem. To solve this two-dimensional problem, a convolutional long short-term memory (ConvLSTM) was applied as the meta-learner, which is shown to be suitable for two-dimensional spatial-temporal data analysis. This model comprised a CNN as the encoder for reading and extracting important features from the input and a vanilla LSTM as the decoder for interpreting the output of the encoder. Several subsequences were needed for each sample in order to fit the model to our univariate time series analysis. Thus, the output sequences of base-learners were employed as these subsequences. The model was constructed of a ConvLSTM2D layer with 64 nodes, followed by a flatten layer to flatten the outputs before being interpreted. The fixed-length output was then provided using a RepeatVector layer, and the output sequence was fed to an LSTM layer with 200 nodes as the input. Next, a Dense layer with 100 nodes was used for interpreting time steps, along with the output layer. A TimeDistributed wrapper was also used to have the prediction for each time step. Further, ReLU, MSE, and Adam were used for all hidden layers as the activation function, loss function, and optimiser, respectively. The model was trained with 500 epochs with a learning rate of 0.01 and an epoch size of 32. Figure 3.4c displays a schematic of the developed method for the prediction horizon of 30 minutes.

3.2.4 Evaluation criteria

The performance of the developed models was evaluated using RMSE and MAE, as regression-wised criteria, and MCC and SE, as clinical-wised criteria described in Section 2.4.1.

3.2.5 Statistical analyses

To statistically compare the performances of all the seven models on the 12 datasets of T1DM data contributors, the non-parametric Friedman test [90] was performed. Then, to pairwisely determine differences, post-hoc analysis utilising Wilcoxon test [89] was done. A significance threshold of 5% was considered. Also, to visualise the post-hoc results, a CD diagram [86] was employed.



(a) The Stacking approach

(b) The Multivariate approach

(c) The Subsequences approach

Figure 3.4: Diagrams of the proposed Stacking approach (a), Multivariate approach (b), and Subsequences approach (c) for the BGL prediction 30 minutes in advance by considering the Linear, VLSTM, and BiLSTM models as base-learners. In the Stacking approach, the output vectors of the base-learners were concatenated and fed as the input to the Linear meta-learner. In Multivariate approach, the output vectors of base-learners were considered as three different variables and fed to a multivariate LSTM meta-learner. In the Subsequences approach, the output vectors of base-leaners were considered as different subsequences for a two-dimensional ConvLSTM encoder-decoder meta-learner.
Note. VLSTM: Vanilla long short-term memory; BiLSTM: Bidirectional long short-term memory; BGL: Blood glucose level; ConvLSTM: convolutional long short-term memory.

3.3 Results and discussion

In this section, the results of all evaluation criteria consisting of RMSE, MAE, MCC, and SE are presented for baseline, non-ensemble, and ensemble models in both horizons of 30 and 60 minutes. The training and testing sets in the Ohio dataset were used for training and evaluation purposes, respectively. The extrapolated data in test sets were excluded in the calculation of evaluation metrics. In addition, due to their stochastic nature, ANN models with performance depending on random initialisation were run five times. The mean and standard deviation of results over the five runs are reported in this section.

3.3.1 Baseline model

Table 3.2 presents the evaluation results for the naive baseline model, which returns the last known value. The results show average evaluation criteria over the 12 patients for both prediction horizons of 30 and 60 minutes.

Table 3.2: Evaluation results of the naive baseline model for prediction horizons of 30 and 60 minutes.

PID	PH: 30 min				PH: 60 min			
	RMSE	MAE	MCC	SE	RMSE	MAE	MCC	SE
540	28.42	21.15	0.67	0.31	47.62	36.19	0.47	0.51
544	22.33	16.47	0.70	0.24	37.56	28.33	0.47	0.40
552	20.62	14.75	0.73	0.23	33.51	24.44	0.55	0.38
559	23.16	16.63	0.75	0.23	39.05	28.74	0.54	0.39
563	20.75	15.44	0.70	0.23	33.95	25.52	0.49	0.36
567	27.37	19.81	0.64	0.30	45.51	33.55	0.38	0.51
570	18.97	13.85	0.83	0.14	31.84	24.26	0.71	0.24
575	25.66	17.83	0.71	0.27	39.83	28.95	0.52	0.44
584	24.64	17.77	0.72	0.24	40.99	29.69	0.54	0.39
588	21.95	16.06	0.72	0.21	35.86	26.74	0.56	0.35
591	24.41	17.96	0.63	0.30	38.37	28.97	0.40	0.48
596	21.03	15.21	0.69	0.24	35.16	25.76	0.48	0.39
Avg	23.27	16.91	0.71	0.25	38.27	28.43	0.51	0.40

Note. PID: Patient ID; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

3.3.2 Non-ensemble models

Tables 3.3 and 3.4 provide the evaluation criteria of the three non-ensemble models for the BGL prediction, 30 and 60 minutes in advance, respectively. Comparing the

results of Tables 3.3 and 3.4 with those in Table 3.2, all developed non-ensemble models outperformed the baseline model regarding all evaluation criteria for both prediction horizons. Considering the VLSTM model, the average of evaluation metrics over all patients for the prediction horizon of 30 minutes was 19.83, 14.09, 0.748, and 0.209 for RMSE, MAE, MCC, and SE, implying an improvement of 14.78%, 16.67%, 5.79%, and 15.04% for these metrics, respectively, compared to the baseline.

Table 3.3: Evaluation results of non-ensemble models for the prediction horizon of 30 minutes.

PID	Model	RMSE	MAE	MCC	SE
	Linear	22.08	16.60	0.74	0.24
540	VLSTM	21.78 ± 0.12	16.25 ± 0.07	0.74 ± 0.00	0.24 ± 0.00
	BiLSTM	22.60 ± 0.78	16.72 ± 0.29	0.72 ± 0.01	0.25 ± 0.00
	Linear	18.10	13.34	0.79	0.20
544	VLSTM	18.09 ± 0.30	13.02 ± 0.22	0.79 ± 0.01	0.19 ± 0.00
	BiLSTM	18.35 ± 1.29	13.29 ± 1.08	0.79 ± 0.02	0.19 ± 0.02
	Linear	16.79	12.77	0.74	0.21
552	VLSTM	16.79 ± 0.09	12.61 ± 0.11	0.75 ± 0.01	0.21 ± 0.00
	BiLSTM	17.16 ± 0.16	12.78 ± 0.14	0.73 ± 0.00	0.21 ± 0.00
	Linear	19.32	13.69	0.80	0.19
559	VLSTM	19.26 ± 0.05	13.52 ± 0.04	0.79 ± 0.01	0.20 ± 0.00
	BiLSTM	20.36 ± 0.67	14.31 ± 0.56	0.78 ± 0.01	0.20 ± 0.01
	Linear	19.25	13.16	0.76	0.18
563	VLSTM	18.94 ± 0.12	13.02 ± 0.03	0.77 ± 0.01	0.18 ± 0.00
	BiLSTM	18.62 ± 0.10	13.03 ± 0.06	0.76 ± 0.01	0.18 ± 0.00
	Linear	21.01	15.13	0.62	0.26
567	VLSTM	20.70 ± 0.06	14.74 ± 0.06	0.66 ± 0.00	0.25 ± 0.00
	BiLSTM	21.48 ± 0.44	15.39 ± 0.34	0.65 ± 0.02	0.26 ± 0.01
	Linear	16.59	11.87	0.86	0.11
570	VLSTM	16.46 ± 0.13	11.43 ± 0.17	0.86 ± 0.00	0.11 ± 0.00
	BiLSTM	16.79 ± 0.64	11.71 ± 0.60	0.86 ± 0.01	0.11 ± 0.01
	Linear	24.35	15.68	0.74	0.24
575	VLSTM	24.20 ± 0.31	15.46 ± 0.09	0.73 ± 0.00	0.24 ± 0.00
	BiLSTM	24.23 ± 0.48	15.81 ± 0.43	0.72 ± 0.01	0.24 ± 0.01

(continued on next page)

Table 3.3 (continued)

PID	Model	RMSE	MAE	MCC	SE
	Linear	21.96	16.10	0.76	0.22
584	VLSTM	22.58 ± 0.19	16.58 ± 0.19	0.76 ± 0.00	0.23 ± 0.00
	BiLSTM	22.05 ± 0.34	16.03 ± 0.36	0.77 ± 0.00	0.22 ± 0.01
	Linear	19.22	14.10	0.75	0.19
588	VLSTM	19.47 ± 0.14	14.11 ± 0.09	0.73 ± 0.00	0.19 ± 0.00
	BiLSTM	19.16 ± 0.14	13.83 ± 0.09	0.74 ± 0.01	0.18 ± 0.00
	Linear	21.74	15.92	0.63	0.27
591	VLSTM	21.82 ± 0.15	15.65 ± 0.10	0.65 ± 0.00	0.27 ± 0.00
	BiLSTM	22.20 ± 0.59	16.12 ± 0.61	0.64 ± 0.01	0.28 ± 0.01
	Linear	17.82	12.81	0.73	0.21
596	VLSTM	17.86 ± 0.09	12.68 ± 0.13	0.75 ± 0.00	0.20 ± 0.00
	BiLSTM	17.57 ± 0.14	12.47 ± 0.13	0.75 ± 0.00	0.20 ± 0.00
	Linear	19.85	14.26	0.74	0.21
Avg	VLSTM	19.83 ± 0.05	14.09 ± 0.04	0.75 ± 0.00	0.21 ± 0.00
	BiLSTM	20.05 ± 0.14	14.29 ± 0.10	0.74 ± 0.00	0.21 ± 0.00

Note. PID: Patient ID; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Table 3.4: Evaluation results of non-ensemble models for the prediction horizon of 60 minutes.

PID	Model	RMSE	MAE	MCC	SE
	Linear	41.10	31.81	0.53	0.44
540	VLSTM	44.94 ± 2.89	33.13 ± 1.27	0.55 ± 0.02	0.44 ± 0.01
	BiLSTM	40.80 ± 0.74	31.22 ± 0.51	0.56 ± 0.01	0.42 ± 0.00
	Linear	31.82	24.68	0.61	0.36
544	VLSTM	31.59 ± 0.46	24.15 ± 0.54	0.60 ± 0.01	0.35 ± 0.01
	BiLSTM	31.30 ± 0.12	24.02 ± 0.13	0.62 ± 0.01	0.35 ± 0.00
	Linear	30.25	23.65	0.59	0.36
552	VLSTM	30.37 ± 0.47	23.34 ± 0.47	0.58 ± 0.01	0.35 ± 0.01
	BiLSTM	30.30 ± 0.13	22.98 ± 0.27	0.58 ± 0.00	0.35 ± 0.00
	Linear	33.73	24.86	0.63	0.34
559					

(continued on next page)

Table 3.4 (continued)

PID	Model	RMSE	MAE	MCC	SE
	VLSTM	35.03 ± 0.74	25.91 ± 0.50	0.62 ± 0.01	0.35 ± 0.01
	BiLSTM	34.00 ± 0.55	24.77 ± 0.31	0.63 ± 0.00	0.34 ± 0.00
	Linear	30.47	22.08	0.56	0.30
563	VLSTM	31.12 ± 0.26	22.38 ± 0.31	0.55 ± 0.01	0.30 ± 0.00
	BiLSTM	30.30 ± 0.24	22.01 ± 0.20	0.56 ± 0.02	0.30 ± 0.00
	Linear	37.56	28.34	0.35	0.47
567	VLSTM	37.39 ± 0.46	28.29 ± 0.40	0.38 ± 0.01	0.47 ± 0.01
	BiLSTM	39.01 ± 1.54	29.42 ± 1.23	0.35 ± 0.04	0.49 ± 0.02
	Linear	28.71	21.41	0.75	0.20
570	VLSTM	28.10 ± 0.41	20.04 ± 0.22	0.78 ± 0.00	0.19 ± 0.00
	BiLSTM	29.23 ± 0.55	21.49 ± 0.58	0.75 ± 0.01	0.20 ± 0.01
	Linear	37.65	27.34	0.53	0.41
575	VLSTM	37.80 ± 0.50	27.08 ± 0.33	0.50 ± 0.01	0.41 ± 0.01
	BiLSTM	37.38 ± 0.34	27.26 ± 0.59	0.50 ± 0.01	0.40 ± 0.01
	Linear	36.64	27.58	0.60	0.37
584	VLSTM	38.09 ± 1.54	28.52 ± 1.34	0.61 ± 0.02	0.38 ± 0.02
	BiLSTM	37.60 ± 0.13	28.22 ± 0.12	0.62 ± 0.00	0.37 ± 0.00
	Linear	31.86	23.48	0.55	0.31
588	VLSTM	31.87 ± 0.08	23.41 ± 0.04	0.54 ± 0.00	0.31 ± 0.00
	BiLSTM	32.08 ± 0.23	23.48 ± 0.28	0.55 ± 0.01	0.31 ± 0.00
	Linear	34.00	26.75	0.40	0.44
591	VLSTM	34.50 ± 0.61	26.67 ± 0.63	0.42 ± 0.02	0.43 ± 0.01
	BiLSTM	34.71 ± 0.33	26.78 ± 0.36	0.43 ± 0.02	0.43 ± 0.01
	Linear	29.72	22.16	0.54	0.33
596	VLSTM	29.77 ± 0.21	21.85 ± 0.16	0.58 ± 0.01	0.33 ± 0.00
	BiLSTM	29.77 ± 0.20	21.90 ± 0.14	0.57 ± 0.01	0.33 ± 0.00
	Linear	33.63	25.34	0.55	0.361
Avg	VLSTM	34.21 ± 0.15	25.40 ± 0.04	0.56 ± 0.00	0.36 ± 0.00
	BiLSTM	33.87 ± 0.12	25.29 ± 0.11	0.56 ± 0.00	0.36 ± 0.00

Note. PID: Patient ID; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Based on the comparison results of the non-ensemble models (Tables 3.3 and

3.4) and Figure 3.3, it can be seen that the performance of the Linear model was considerably better than the two LSTM models in the tuning process. However, this difference was negligible in the final evaluation process. This deviation is plausible because, in the final evaluation, a larger dataset for training was used —in tuning, 80% of the training data were used for training purposes rather than all. It is postulated that more training data can improve the performance of the deep learning models [157].

3.3.3 Ensemble models

The evaluation results of the three developed ensemble models for both prediction horizons of 30 and 60 minutes are listed in Tables 3.5 and 3.5, respectively.

It is notable to feed a unique input to meta-learners, among the five non-ensemble VLSTM and BiLSTM trained models, the model with the lowest RMSE on the 20% of the training data allocated to the validation data was selected for each base-learner. Then, the ensemble models were run five times, and the mean and standard deviation over the five runs are presented accordingly.

Table 3.5: Evaluation results of ensemble models for the prediction horizon of 30 minutes.

PID	Model	RMSE	MAE	MCC	SE
540	Stacking	21.98	16.16	0.74	0.24
	Multivariate	21.46 ± 0.07	16.11 ± 0.04	0.73 ± 0.00	0.24 ± 0.00
	Subsequences	21.54 ± 0.18	16.06 ± 0.05	0.73 ± 0.00	0.24 ± 0.00
544	Stacking	17.83	12.73	0.79	0.18
	Multivariate	17.88 ± 0.04	12.79 ± 0.04	0.79 ± 0.00	0.18 ± 0.00
	Subsequences	17.92 ± 0.10	12.80 ± 0.08	0.79 ± 0.00	0.18 ± 0.00
552	Stacking	16.42	12.13	0.76	0.120
	Multivariate	16.70 ± 0.03	12.48 ± 0.03	0.74 ± 0.00	0.20 ± 0.00
	Subsequences	16.68 ± 0.03	12.41 ± 0.04	0.74 ± 0.00	0.20 ± 0.00
559	Stacking	19.33	13.37	0.79	0.20
	Multivariate	19.45 ± 0.26	13.47 ± 0.09	0.79 ± 0.00	0.19 ± 0.00
	Subsequences	19.27 ± 0.12	13.33 ± 0.09	0.79 ± 0.00	0.19 ± 0.00
563	Stacking	18.86	12.97	0.77	0.18
	Multivariate	18.61 ± 0.05	12.91 ± 0.04	0.77 ± 0.00	0.18 ± 0.00
	Subsequences	18.56 ± 0.10	12.88 ± 0.03	0.77 ± 0.00	0.18 ± 0.00

(continued on next page)

Table 3.5 (continued)

PID	Model	RMSE	MAE	MCC	SE
	Stacking	20.49	14.55	0.68	0.24
567	Multivariate	20.52 ± 0.04	14.60 ± 0.06	0.67 ± 0.01	0.24 ± 0.00
	Subsequences	20.59 ± 0.07	14.67 ± 0.04	0.67 ± 0.00	0.25 ± 0.00
	Stacking	16.39	11.24	0.87	0.11
570	Multivariate	16.48 ± 0.09	11.37 ± 0.13	0.86 ± 0.00	0.11 ± 0.00
	Subsequences	16.44 ± 0.06	11.30 ± 0.06	0.86 ± 0.00	0.11 ± 0.00
	Stacking	23.38	15.25	0.74	0.23
575	Multivariate	23.86 ± 0.08	15.39 ± 0.01	0.73 ± 0.00	0.23 ± 0.00
	Subsequences	23.89 ± 0.12	15.38 ± 0.06	0.73 ± 0.00	0.23 ± 0.00
	Stacking	22.08	16.01	0.77	0.22
584	Multivariate	21.89 ± 0.12	15.85 ± 0.13	0.76 ± 0.01	0.22 ± 0.00
	Subsequences	21.97 ± 0.13	15.97 ± 0.17	0.76 ± 0.00	0.22 ± 0.00
	Stacking	19.60	14.08	0.75	0.18
588	Multivariate	19.41 ± 0.11	14.00 ± 0.11	0.74 ± 0.00	0.18 ± 0.00
	Subsequences	19.20 ± 0.10	13.85 ± 0.10	0.75 ± 0.00	0.18 ± 0.00
	Stacking	21.50	15.64	0.64	0.27
591	Multivariate	21.78 ± 0.09	15.62 ± 0.05	0.66 ± 0.00	0.27 ± 0.00
	Subsequences	21.75 ± 0.05	15.57 ± 0.05	0.65 ± 0.00	0.27 ± 0.00
	Stacking	17.70	12.39	0.76	0.20
596	Multivariate	17.70 ± 0.06	12.42 ± 0.04	0.75 ± 0.00	0.20 ± 0.00
	Subsequences	17.63 ± 0.18	12.34 ± 0.08	0.76 ± 0.00	0.20 ± 0.00
	Stacking	19.63	13.88	0.76	0.20
Avg	Multivariate	19.64 ± 0.02	13.92 ± 0.01	0.75 ± 0.00	0.20 ± 0.00
	Subsequences	19.62 ± 0.02	13.88 ± 0.01	0.75 ± 0.00	0.20 ± 0.00

Note. PID: Patient ID; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Table 3.6: Evaluation results of ensemble models for the prediction horizon of 60 minutes.

PID	Model	RMSE	MAE	MCC	SE
	Stacking	40.43	30.64	0.56	0.41
540	Multivariate	40.25 ± 0.13	30.81 ± 0.06	0.56 ± 0.00	0.42 ± 0.00

(continued on next page)

Table 3.6 (continued)

PID	Model	RMSE	MAE	MCC	SE
	Subsequences	40.25 ± 0.42	30.62 ± 0.28	0.57 ± 0.01	0.42 ± 0.00
	Stacking	30.82	22.87	0.63	0.32
544	Multivariate	30.96 ± 0.07	23.23 ± 0.14	0.62 ± 0.00	0.33 ± 0.00
	Subsequences	31.07 ± 0.19	23.16 ± 0.11	0.62 ± 0.00	0.33 ± 0.00
	Stacking	30.24	22.54	0.60	0.34
552	Multivariate	30.02 ± 0.06	22.84 ± 0.09	0.58 ± 0.00	0.35 ± 0.00
	Subsequences	29.95 ± 0.13	22.57 ± 0.21	0.58 ± 0.01	0.35 ± 0.00
	Stacking	35.10	25.55	0.65	0.34
559	Multivariate	34.91 ± 0.18	25.48 ± 0.08	0.63 ± 0.00	0.34 ± 0.00
	Subsequences	34.95 ± 0.16	25.47 ± 0.09	0.63 ± 0.00	0.34 ± 0.00
	Stacking	30.92	22.02	0.57	0.30
563	Multivariate	30.91 ± 0.32	22.12 ± 0.16	0.56 ± 0.00	0.30 ± 0.00
	Subsequences	30.69 ± 0.27	22.08 ± 0.13	0.56 ± 0.01	0.30 ± 0.00
	Stacking	36.51	27.69	0.38	0.46
567	Multivariate	37.06 ± 0.10	27.78 ± 0.07	0.38 ± 0.00	0.46 ± 0.00
	Subsequences	37.52 ± 0.65	27.93 ± 0.35	0.39 ± 0.01	0.46 ± 0.00
	Stacking	27.63	19.93	0.78	0.19
570	Multivariate	27.94 ± 0.12	20.16 ± 0.08	0.78 ± 0.00	0.19 ± 0.00
	Subsequences	28.01 ± 0.25	20.23 ± 0.24	0.77 ± 0.01	0.19 ± 0.00
	Stacking	37.01	26.40	0.52	0.40
575	Multivariate	37.40 ± 0.23	26.63 ± 0.11	0.50 ± 0.01	0.40 ± 0.00
	Subsequences	36.88 ± 0.74	25.98 ± 0.41	0.51 ± 0.01	0.39 ± 0.01
	Stacking	36.92	27.59	0.62	0.36
584	Multivariate	37.14 ± 0.26	27.40 ± 0.34	0.61 ± 0.01	0.36 ± 0.01
	Subsequences	37.15 ± 0.15	27.39 ± 0.27	0.62 ± 0.00	0.36 ± 0.01
	Stacking	31.77	23.18	0.56	0.30
588	Multivariate	31.90 ± 0.19	23.35 ± 0.15	0.55 ± 0.00	0.31 ± 0.00
	Subsequences	31.90 ± 0.07	23.39 ± 0.07	0.55 ± 0.00	0.31 ± 0.00
	Stacking	33.87	25.65	0.44	0.42
591	Multivariate	34.01 ± 0.19	26.06 ± 0.10	0.43 ± 0.00	0.42 ± 0.00
	Subsequences	34.17 ± 0.28	26.03 ± 0.15	0.45 ± 0.01	0.42 ± 0.00
	Stacking	30.19	21.77	0.58	0.32
596	Multivariate	30.37 ± 0.28	21.97 ± 0.12	0.59 ± 0.00	0.32 ± 0.00

(continued on next page)

Table 3.6 (continued)

PID	Model	RMSE	MAE	MCC	SE
	Subsequences	30.80 ± 0.35	22.15 ± 0.16	0.59 ± 0.00	0.32 ± 0.00
Avg	Stacking	33.45	24.65	0.57	0.35
	Multivariate	33.57 ± 0.03	24.82 ± 0.03	0.57 ± 0.00	0.35 ± 0.00
	Subsequences	33.61 ± 0.04	24.75 ± 0.06	0.57 ± 0.00	0.35 ± 0.00

Note. PID: Patient ID; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

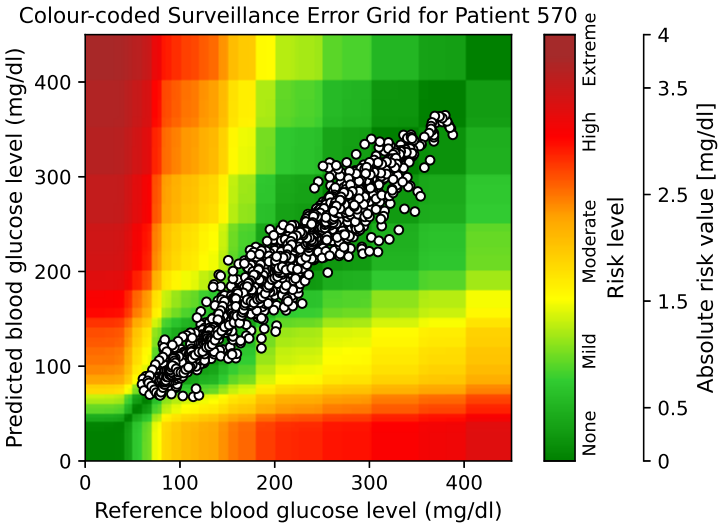
According to the comparison of results in Tables 3.2, 3.5, and 3.6, all developed ensemble models performed better than the baseline regarding all evaluation criteria for both prediction horizons. Considering the Stacking model among ensemble models, the average values of evaluation metrics over all patients for the prediction horizon of 30 minutes were 19.63, 13.88, 0.756, and 0.204 for RMSE, MAE, MCC, and SE, indicating an improvement of 15.64%, 17.91%, 6.93%, and 17.07% for these metrics, respectively, in comparison with the baseline. This model also made an improvement of 12.59%, 13.29%, 12.96%, and 13.86% for RMSE, MAE, MCC, and SE metrics for the prediction horizon of 60 minutes, respectively.

According to the comparison between the results of Tables 3.3, 3.4, 3.5, and 3.6, ensemble models outperformed non-ensemble models for both prediction horizons. Further, it is worth mentioning that these improvements happened while due to computational costs, the meta-learners of ensemble models were not fine-tuned, but the hyperparameter optimisation was performed for non-ensemble models.

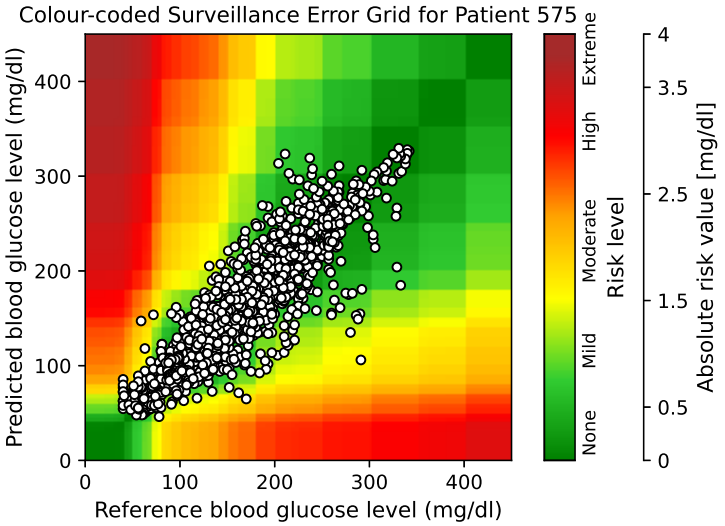
The colour-coded SEGs related to the predictions of the Stacking model 30 minutes in advance for patients 570 and 575 (with the best and the worst evaluation results, respectively) are illustrated in Figure 3.5 to have a clinical insight into BGL predictions. As shown in Figure 3.5a, BGL predictions for patient 570 are in the none and mild risk regions. However, some predictions are placed in the moderate to high risk regions for patient 575 in Figure 3.5b.

3.3.4 Statistical analyses

According to each evaluation metric, the CD diagrams, where a thick horizontal line connects groups of not-significantly different prediction models, are presented. Figures 3.6 and 3.7 show CD diagrams related to the comparison of all prediction models against each other over the 12 datasets of T1DM data contributors according to each evaluation metric for prediction horizons of 30 and 60 minutes, respectively.



(a)



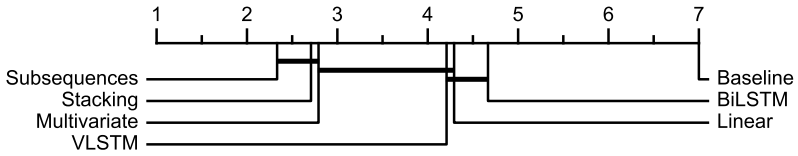
(b)

Figure 3.5: The colour-coded surveillance error grid of the Stacking approach for patients 570 (a) and 575 (b). The white circles illustrate blood glucose level predictions and the corresponding reference blood glucose levels. In addition, the risk value of each prediction comparing with its reference value was coded by colour. There are five categories for a risk level, including none, mild, moderate, high, and extreme.

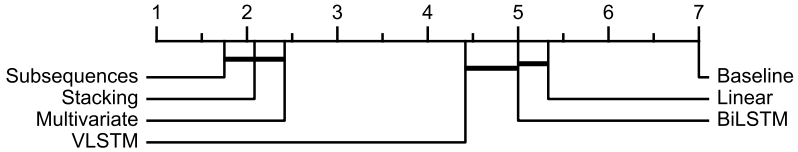
Moreover, to have an statistical overview, Figure 3.8 graphically represents CD diagrams according to the average ranking over all evaluation criteria (RMSE, MAE, MCC, and SE) for both prediction horizons of 30 (3.8a) and 60 (3.8b) minutes.

Considering the statistical analyses, it can be concluded that three non-ensemble

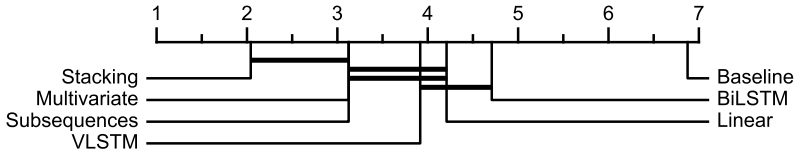
models predicted BGL with the statistically significant improvement compared with the baseline model and no overall significant difference in between. Also, the ensemble models performed statistically significantly better than baseline and non-ensemble models with no significant intra-difference.



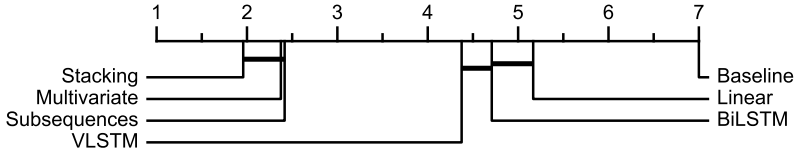
(a) Metric: RMSE, PH: 30 min



(b) Metric: MAE, PH: 30 min



(c) Metric: MCC, PH: 30 min

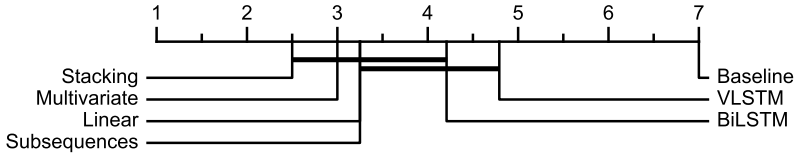


(d) Metric: SE, PH: 30 min

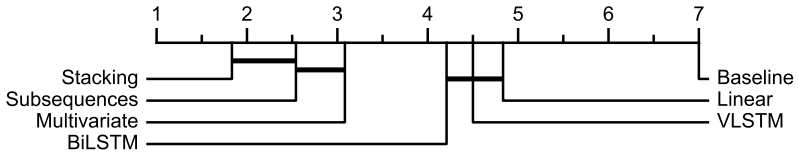
Figure 3.6: Critical difference diagram showing comparison of all prediction models against each other over the 12 datasets of T1DM data contributors according to RMSE (a), MAE (b), MCC (c), and SE (d) for prediction horizon of 30 minutes.

3.3.5 Computational analysis

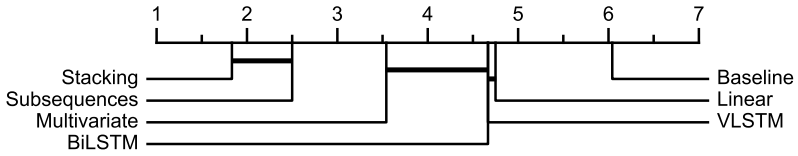
The developed models rely on exploiting patterns in BGL data for the prediction. Therefore, changes in the patterns, for example, when a person's habit changes, may require a readjustment to the prediction models. Hence, it is valuable to investigate the time for retraining models relative to the time required for new data collection. The average execution time of training the developed models across all patients for running codes using a commodity laptop computer (specifications: core i7 2.8 GHz



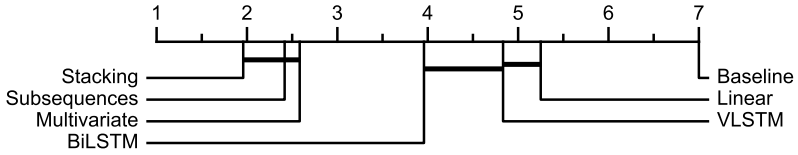
(a) Metric: RMSE, PH: 60 min



(b) Metric: MAE, PH: 60 min



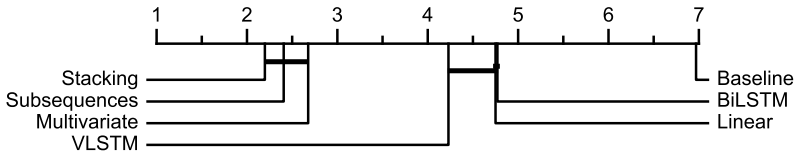
(c) Metric: MCC, PH: 60 min



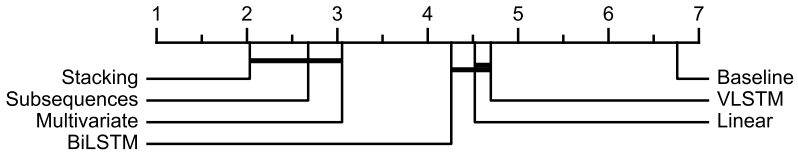
(d) Metric: SE, PH: 60 min

Figure 3.7: Critical difference diagram showing comparison of all prediction models against each other over the 12 datasets of T1DM data contributors according to RMSE (a), MAE (b), MCC (c), and SE (d) for prediction horizon of 60 minutes.

processor, 16 GB of RAM, and NVIDIA GeForce GTX 1050 Ti GPU) approximately was: a few seconds for the baseline and Linear models, 40 minutes for the VLSTM, 50 minutes for BiLSTM, 90 minutes for the Stacking, 120 minutes for the Multivariate, and 170 minutes for the Subsequences. Although the training times of developed ensemble models are considerably longer than the non-ensemble models, these training times are considerably less than the time required for collecting new data for retraining purposes. Also, it is worth remarking that the simple Linear model produced results comparable to the two more complicated LSTM models, which are popular for time series forecasting. It could imply that even a slight improvement in the BGL prediction task would be challenging, and it could not be an easy trade-off between the complexity and accuracy of the prediction. Hence, a slight improvement in ensemble approaches could be appreciable.



(a) PH: 30min



(b) PH: 60min

Figure 3.8: Critical difference diagram showing comparison of all prediction models against each other over the 12 datasets of T1DM data contributors according to average over all criteria for prediction horizon of 30 (a) and 60 (b) minutes.

3.4 Summary

This work contributes to predicting BGL 30 and 60 minutes in advance by proposing three methodologies using deep and ensemble learning and comparing their performance with three non-ensemble benchmark models as well as a naive baseline model. The Linear, VLSTM, and BiLSTM models were the applied non-ensemble models. The benchmark models were used as base-learners for developing the ensemble models. The outputs of the base-learners were then fused using the meta-learning approach in three different ways, including univariate time series forecasting, multivariate time series forecasting, and two-dimensional data analysis. The relevant resultant ensemble models were named Stacking, Multivariate, and Subsequences, respectively. In the Stacking approach, the output vectors of the base-learners were concatenated and fed to the Linear model as the meta-learner. In the Multivariate approach, the output vectors of base-learners were considered as different variables. Therefore, the univariate time series forecasting was converted to a multivariate time series analysis using a multivariate LSTM as the meta-learner. In the Subsequences approach, the output vectors of base-learners were considered as different subsequences. The one-dimensional time series forecasting was configured as a two-dimensional data analysis using a ConvLSTM as the meta-learner.

The results obtained show that all the developed non-ensemble models outperformed the naive baseline model. Moreover, the novel advanced ensemble models resulted in a statistically significant improvement over the non-ensemble models. Among all developed ensemble models, the Stacking approach represented slightly

better performance. This work also offered an overview of the feasibility and usefulness of meta-learning in changing the dimension of a univariate time series forecasting task by proposing two novel Multivariate and Subsequences meta-learning approaches which provided results comparable to the Stacking approach.

Chapter 4

Leveraging causal analysis in blood glucose level prediction

4.1 Preface

In causal analysis, the relations between causes and effects are examined. Causal inference, as the main approach in causality analysis, quantifies causal relations [120, 121]. Data-driven causal inference techniques have developed in recent years to discover causal associations between variables in time series data [122]. These approaches assess causality from multivariate time series based on observations of variables in complex systems [123, 124]. Clinically, some factors, including carbohydrates, bolus, and PA, have been identified as the main factors affecting the BGL [14]. Since there are cause and effect relations between these variables and BGL, these relations could be examined in a causality framework. Furthermore, causality has recently been applied in time series forecasting tasks in several domains, such as neuroscience [33], climate [35], and economic data [132, 133]. Using causality-based influencing features, these applications attempt to improve the performance of time series forecasting. However, in fields like predicting BGL in diabetes, where the affecting variables on BGL are known and limited, causality cannot be used in the same way. Hence, there is a lack of causality-based approaches for time series forecasting in the field of diabetes management.

The work presented in this chapter initially investigates the relations between BGL and affecting variables, including carbohydrates, bolus, and PA, and quantifies the causality strengths of each variable with BGL using the convergent cross mapping (CCM) method, as an appropriate approach for causality investigation of complex systems. Then, considering the extended convergent cross mapping (ECCM) method, the causality strengths of each variable for different lags are quantified.

After that, the optimal time lag for each variable is determined according to the quantified causality effects. Subsequently, the feasibility of leveraging causality information as prior knowledge for BGL prediction is investigated by proposing two approaches. In the first approach, causality strengths are used as weights for relevant affecting variables. In the second approach, the optimal causal lags and the corresponding causality strengths are considered the shifts and weights for the variables, respectively. Evaluation and statistical analyses are then used to assess and compare the performance of BGL prediction with and without deploying causality to validate the impact of causality usage in BGL prediction performance.

4.2 Material and methods

4.2.1 Dataset

The work presented in this chapter also used the Ohio T1DM datasets [66, 67] explained in Section 2.8. In this work, we used data related to BGL, carbohydrate, bolus, and PA (HR for the Ohio_2018 dataset and MA for the Ohio_2020 dataset).

4.2.2 Preprocessing

The first step in the preprocessing was dealing with the missing data. Table 4.1 presents the number and percentage of missing data points for BGL, HR, and MA in training and testing sets related to the contributors in both cohorts. The missing data were imputed linearly. To accomplish this, training sets were interpolated with linear interpolation. Also, missing values in the testing sets were imputed using linear extrapolation to guarantee that future data would not be observed by the model and that it could be applied in a real-time context.

The next step was to unify the resolution of BGL and other data for alignment. For this purpose, MA data was downsampled to 5-minute intervals by capturing the nearest recorded data with each BGL data and excluding the remainder. Also, due to the difference in wear time of CGM and activity sensors, there were a considerable number of unavailable activity and/or BGL data at the beginning and/or end of each set. These timestamps were discarded for the alignment to have a more reliable analysis. Moreover, data of carbohydrate and bolus were inputted with zero for timestamps with no reported data.

For the BGL prediction, we dealt with the time series forecasting task. The time series problem was reframed to a supervised learning task. For this purpose, the time series data were transformed into samples with lag observations as input and

Table 4.1: The number and percentage of missing data points for BGL and activity in training and testing sets related to the contributors in the Ohio_2018 and Ohio_2020 datasets.

	PID	Training		Testing	
		BG	HR	BG	HR
Ohio_2018	559	1285 (10.64%)	451 (3.78%)	362 (12.59%)	175 (6.23%)
	563	974 (7.44%)	433 (3.61%)	121 (4.50%)	174 (6.04%)
	570	629 (5.42%)	275 (2.36%)	135 (4.69%)	88 (3.13%)
	575	1238 (9.45%)	965 (7.36%)	128 (4.71%)	60 (2.18%)
	588	466 (3.56%)	392 (2.99%)	89 (3.09%)	114 (4.23%)
	591	1908 (14.96%)	1233 (9.65%)	87 (3.06%)	212 (7.36%)
	PID	BG	MA	BG	MA
Ohio_2020	540	1163 (8.87%)	7087 (54.11%)	170 (5.54%)	953 (37.82%)
	544	2049 (16.17%)	1385 (11.04%)	420 (13.39%)	297 (9.38%)
	552	2017 (18.18%)	5427 (51.23%)	1586 (40.15%)	1214 (36.47%)
	567	2678 (19.78%)	5259 (39.16%)	482 (16.79%)	970 (32.56%)
	584	1098 (8.29%)	3658 (27.87%)	330 (11.02%)	427 (14.27%)
	596	2752 (20.19%)	6198 (43.78%)	260 (8.66%)	1053 (35.62%)

Note. PID: Patient identity; BGL: Blood glucose level; HR: Heart rate; MA: Magnitude of acceleration.

future observations as output. Next, using a rolling window of 5-minutes (corresponding to one data point), samples were created with 90-minute history lengths (corresponding to 18 data points) as input. The associated output of each sample was a vector related to 30 and 60 minutes (corresponding to six and 12 data points, respectively). Next, for each variable, the input sequence was scaled to the minimum and maximum value over the entire training set of that variable.

4.2.3 Causality analysis

Causal discovery and causal inference are two main fields for causality analysis. The former qualitatively investigates causal relationships between variables and estimates a causal network from observational data. The latter examines causal relations quantitatively and estimates the causal effect [124].

In this work, CCM was used to investigate the causal relation between BGL and carbohydrate, bolus, HR, and MA variables. CCM [129] is an appropriate method for the causality investigation of complex systems. Then, the ECCM [158] was used to examine causality strength for different lags up to six hours and to find the effective lag of causation for each variable on BGL.

4.2.3.1 Convergent cross mapping

CCM is from Takens' Theorem, in which if x impacts y, the history of x can be restored from y, which can be achieved using the cross map technique. Accordingly, the capability of x estimation determines the information about x, which is embedded into y. Thus, how well y cross maps x measures the causal effect of x on y. In addition, convergence in CCM means the longer the history length, the better the cross map skill. With convergence, the skill of cross map x by y can qualify the causation [129].

In this work, the relations between carbohydrates, bolus, HR, and MA (as causes) and BGL (as effect) were examined using the CCM method. The causal relation of each variable with BGL was inferred by computing the cross map skill with increasing time series length. The cross map skill was measured using the Pearson correlation with a threshold of five percent as the significance level. Increasing daily to the length of the training set, the lengths were examined. Also, for each length, 10 sampled time series were considered. The embedding lag step and embedding dimension were fixed to one and two, respectively.

4.2.3.2 Extended convergent cross mapping

ECCM [158], as the extension of CCM, proposed investigating different time lags for calculating cross map skills. In this work, the optimal time lag for the impact of each variable on BGL was determined by applying the ECCM. The cross map skill was measured for variables for different time lags up to six hours (corresponding to 72 data points). The optimal time lag was related to the highest value for cross map skill. Also, to speed up computation, the distance between investigated lags was gradually increased every hour. For each lag, average values of cross map skill were examined over 10 sampled time series.

4.2.4 Leveraging causality in BGL prediction

After investigating the causation between variables of T1DM management, we examined the application of causality information in BGL prediction. To do so, we implemented two multivariate prediction models, then, proposed two approaches for deploying causality information as prior knowledge for BGL prediction. Lastly, causality-integrated BGL prediction was evaluated and statistically compared with normal prediction. Brief descriptions for each of these steps are presented in the following.

4.2.4.1 Prediction models

It is important to investigate the performance of approaches with the state-of-the-art models for BGL prediction in T1DM. Although it is difficult to make a fair comparison due to the use of different datasets, inputs, and models [39], in this work, two multivariate prediction models are examined. One model is an LSTM, similar to the model developed in Section 3.2.3.2. The other model is a convolutional recurrent neural network (CRNN), similar to the model proposed by Freiburghaus et al [136]. This model has achieved the first rank in the BGL prediction challenge in 2020. Each model is briefly described in the following.

LSTM The LSTM model was built using a 200-unit LSTM layer, a 100-unit Dense layer, and an output layer with the output points as the number of units. He uniform and ReLU were used for initialiser and activation functions purposes, respectively. Also, Adam and MSE were used as the optimiser and the loss function, respectively. The epoch size and batch size were considered as 200 and 32, respectively. The learning rate was reduced using the ReduceLROnPlateau callback with an initial learning rate of 0.01. The learning rate was decreased by 0.1 with a patient number of 20 epochs when improvement of validation loss had stopped. A history length of 90 minutes was selected via an optimisation process investigating up to three hours of history using the validation subset, which was the last 20% of the training set.

CRNN To implement the CRNN architecture, as mentioned in the original work [136], along with the preprocessing steps described previously, features were smoothed over a window of two hours' worth of history data using a 1D Gaussian filter. The model was built using Convolution 1D, Maxpooling 1D, LSTM, and Dense layers. Also, RMSProp and mean absolute error were used as the optimiser and the loss function, respectively. The epoch size and batch size were considered as 1000 and 1024, respectively. The learning rate was reduced using the ReduceLROnPlateau callback with an initial learning rate of 0.001, a factor of 0.1, and patience of three epochs. Moreover, EarlyStopping callback was used to stop training the model after 50 epochs of stopping validation loss improvement. The details about the architecture of the CRNN can be found in [136]. It is worth clarifying that the mentioned work used different features.

4.2.4.2 Causality knowledge

Two approaches were developed to leverage the causality information captured from CCM and ECCM methods in BGL prediction:

Convergent cross mapping based approach In the convergent cross mapping based approach (CCMBA), causal strength values of variables, quantified using the CCM method, were considered for weighting the corresponding variables before using them as input for the BGL prediction model (Figure 4.1a). The basic idea of this approach is that each variable with stronger causal strength could be more helpful in the prediction.

Extended convergent cross mapping based approach In the extended convergent cross mapping based approach (ECCMBA), the optimal time lag, corresponding to the maximum causal strength, for each variable was determined using the ECCM method. The obtained optimal time lag and the maximum casual strength were then used as a shift and weight for the original variable, respectively (Figure 4.1b). The basic idea of considering optimal lags is that by applying the shifts, more informative histories of affecting variables might be selected.

4.2.5 Evaluation criteria

The performance of BGL prediction using different models and approaches was evaluated using regression-wised criteria, including RMSE and MAE, and clinical-wised criteria, including MCC and SE, as presented in Section 2.4.1.

4.2.6 Statistical analyses

The impact of deploying causality information as prior knowledge in the performance of the BGL prediction model was statistically investigated. The non-parametric Wilcoxon test [89] is a suitable tool for comparison of a newly proposed method with the existing one over multiple datasets with no assumption of normal distribution [86]. Hence, this test was performed to determine if the calculated evaluation criteria for each proposed approach and the Normal approach on the same set of data providers are consistent.

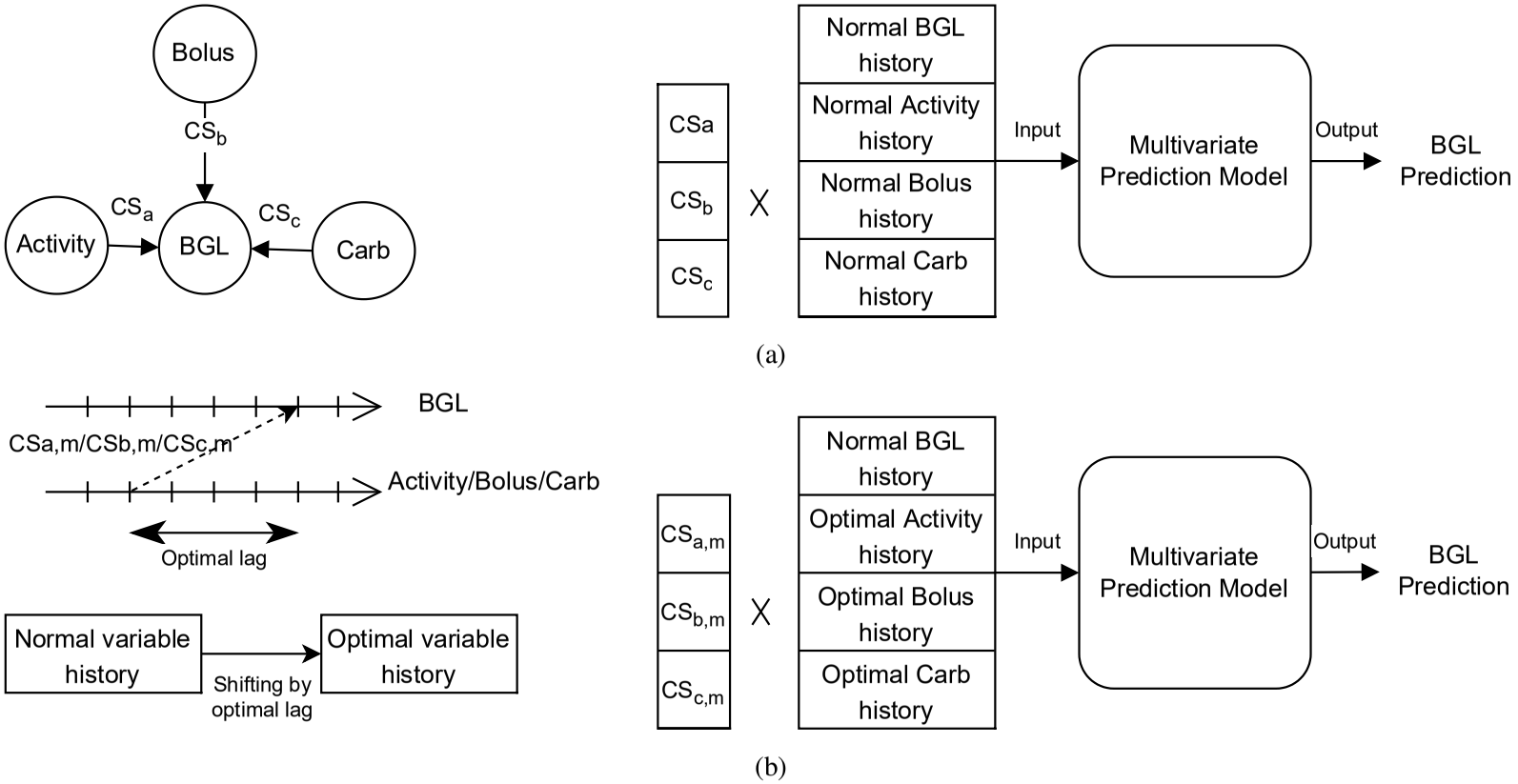


Figure 4.1: Diagrams of the proposed CCMBA (a) and ECCMBA (b) approaches for leveraging causality information as prior knowledge in BGL prediction.

Note. BGL: Blood glucose level; Activity: Physical activity (heart rate or magnitude of acceleration); Bolus: Injected bolus insulin; Carb: Carbohydrate intake; CS_a: Causality strength of activity on BGL; CS_b: Causality strength of Bolus on BGL; CS_c: Causality strength of carbohydrate on BGL; $CS_{m,a}$: Maximum causality strength of activity on BGL; $CS_{m,b}$: Maximum causality strength of Bolus on BGL; $CS_{m,c}$: Maximum causality strength of carbohydrate on BGL.

4.3 Results and discussion

In this section, the results of causality analyses and performance evaluation of BGL prediction with and without leveraging causality information as prior knowledge are presented and discussed. Training sets were used for the causality analysis and training of the BGL prediction model. Also, an evaluation of the BGL prediction performance was performed using testing sets. It is of note that the extrapolated data in testing sets were excluded in the calculation of evaluation metrics so that the assessment was performed only on real data.

4.3.1 Causality analysis

The results of the causality investigation, including causal strengths and optimal time lags of the variables for each data contributor, are presented in this section.

4.3.1.1 CCM

Figures 4.2 and 4.3 show the cross map skill as a function of time series length for patients in Ohio_2018 and Ohio_2020 datasets, respectively. The causality strengths of different variables (carbohydrate, Bolus, and HR for Ohio_2018 dataset, and carbohydrate, bolus, and MA for Ohio_2020 dataset) show similar overall patterns for all patients, where convergences happen with increasing time series length. The non-zero cross map skill values, along with the convergence for the variables, indicate the existence of causation between all the variables and BGL. Also, the corresponding cross map skill values can quantify the causal strength of the variables with BGL.

Considering Figure 4.2, for patients in Ohio_2018, the causality strength of activity, measured as HR, is stronger than that of carbohydrate and bolus. The superiority of casual strength for activity is less considerable in patients with PIDs 559 and 563 (Figures 4.2a and 4.2b) compared to others. Also, with a small difference, the causality strength of carbohydrate is stronger than that of bolus for all patients.

Considering Figure 4.3, for the majority of patients in Ohio_2020, causality strength of activity (MA) is again stronger than that of carbohydrate and bolus. However, the difference between the causality strength of carbohydrates and bolus is not conclusive.

Table 4.2 summarises the mean and standard deviation values over 10 sampled time series related to causality strengths of the variables with BGL for each patient in Ohio_2018 and Ohio_2020 datasets. The average causation strength of each variable over all patients in each dataset is also presented in the table. In both

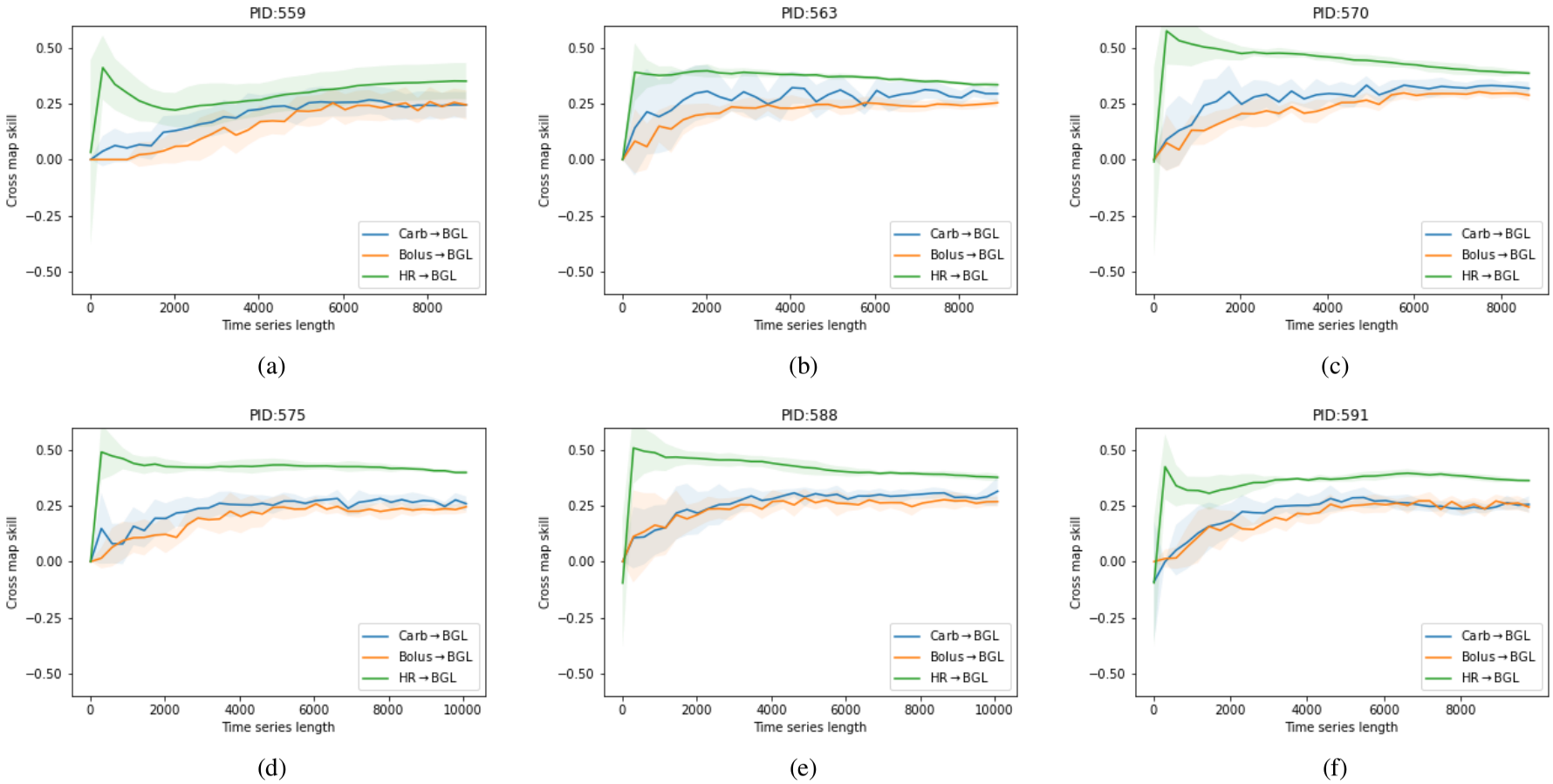


Figure 4.2: The cross map skill as a function of length of time series for PID (a) 559, (b) 563, (c) 570, (d) 575, (e) 588, and (f) 591, in Ohio_2018 dataset. Lines and shaded regions respectively show average and standard deviation over 10 sampled time series.

Note. PID: Patient identity; BGL: Blood glucose level; HR: Heart rate; Bolus: Injected bolus insulin; Carb: Carbohydrate intake.

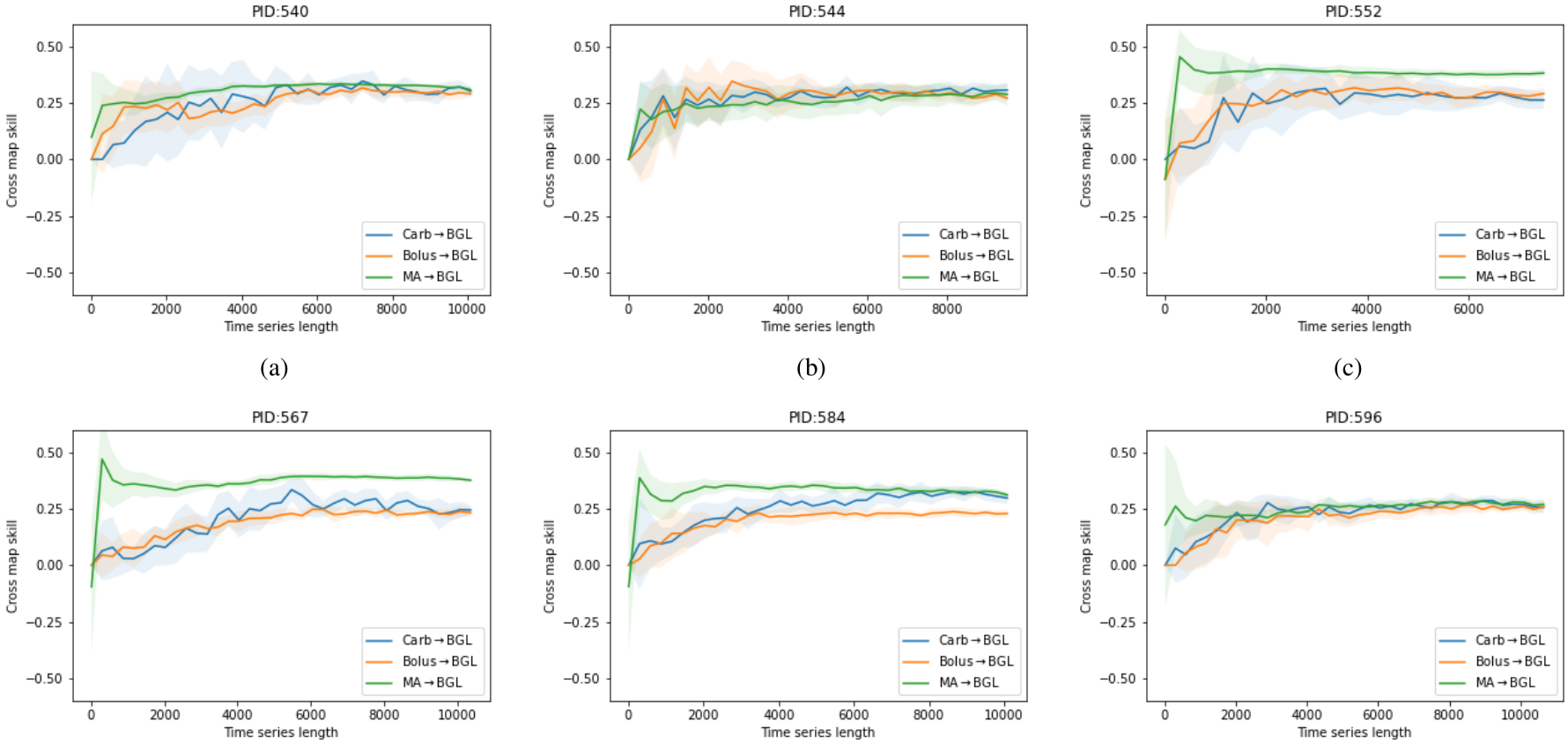


Figure 4.3: The cross map skill as a function of length of time series for PID (a) 540, (b) 544, (c) 552, (d) 567, (e) 584, and (f) 596, for Ohio_2020 dataset. Lines and shaded regions show means and standard deviations over 10 sampled time series.

Note. PID: Patient identity; BGL: Blood glucose level; MA: Magnitude of Acceleration; Bolus: Injected bolus insulin; Carb: Carbohydrate intake.

datasets, comparing the average causal strengths of variables shows that the causation strength of activity (HR in the ohio_2018 dataset and MA in the Ohio_2020 dataset) is higher than that of carbohydrate and bolus. Also, carbohydrate has stronger causation than bolus, comparably.

Table 4.2: The results of causality strength using CCM in Ohio_2018 and Ohio_2020 datasets.

	PID	Carb	Bolus	HR
Ohio_2018	559	0.254 ± 0.020	0.278 ± 0.018	0.375 ± 0.023
	563	0.287 ± 0.028	0.249 ± 0.020	0.330 ± 0.009
	570	0.316 ± 0.024	0.291 ± 0.025	0.383 ± 0.011
	575	0.258 ± 0.018	0.228 ± 0.027	0.393 ± 0.010
	588	0.309 ± 0.073	0.268 ± 0.021	0.376 ± 0.013
	591	0.243 ± 0.029	0.235 ± 0.029	0.359 ± 0.009
	Avg	0.278 ± 0.032	0.258 ± 0.023	0.369 ± 0.013
	PID	Carb	Bolus	MA
Ohio_2020	540	0.283 ± 0.038	0.297 ± 0.019	0.315 ± 0.007
	544	0.302 ± 0.018	0.274 ± 0.021	0.293 ± 0.036
	552	0.285 ± 0.031	0.288 ± 0.029	0.380 ± 0.017
	567	0.244 ± 0.025	0.230 ± 0.018	0.377 ± 0.012
	584	0.312 ± 0.024	0.226 ± 0.016	0.314 ± 0.017
	596	0.262 ± 0.021	0.258 ± 0.025	0.273 ± 0.018
	Avg	0.281 ± 0.026	0.262 ± 0.021	0.325 ± 0.018

Note. PID: Patient identity; BGL: Blood glucose level; Carb: Carbohydrate intake; Bolus: Injected bolus insulin; HR: Heart rate; MA: Magnitude of acceleration.

4.3.1.2 ECCM

Figures 4.4 and 4.5 illustrate the cross map skill as a function of cross map lag for contributors in Ohio_2018 and Ohio_2020 datasets, respectively. For each variable, optimal lag was defined as the highest cross map skill among investigated lags up to six hours. The figures show stronger causality of activity with BGL, compared to carbohydrate and bolus, for different lags for all patients in Ohio_2018 and most patients in Ohio_2020. In comparison, the causal strengths of the carbohydrate and bolus for the examined lags are comparable in both datasets. Although the results do not reveal definitive trends for each variable across all patients, there is a general descending pattern between cross-map skill of HR and time lag in PIDs 559, 575, and 588 (Figures 4.4a, 4.4d, and 4.4e).

The optimal lags and related cross map skill values for the variables are summarised in Table 4.3 for all patients in both datasets.

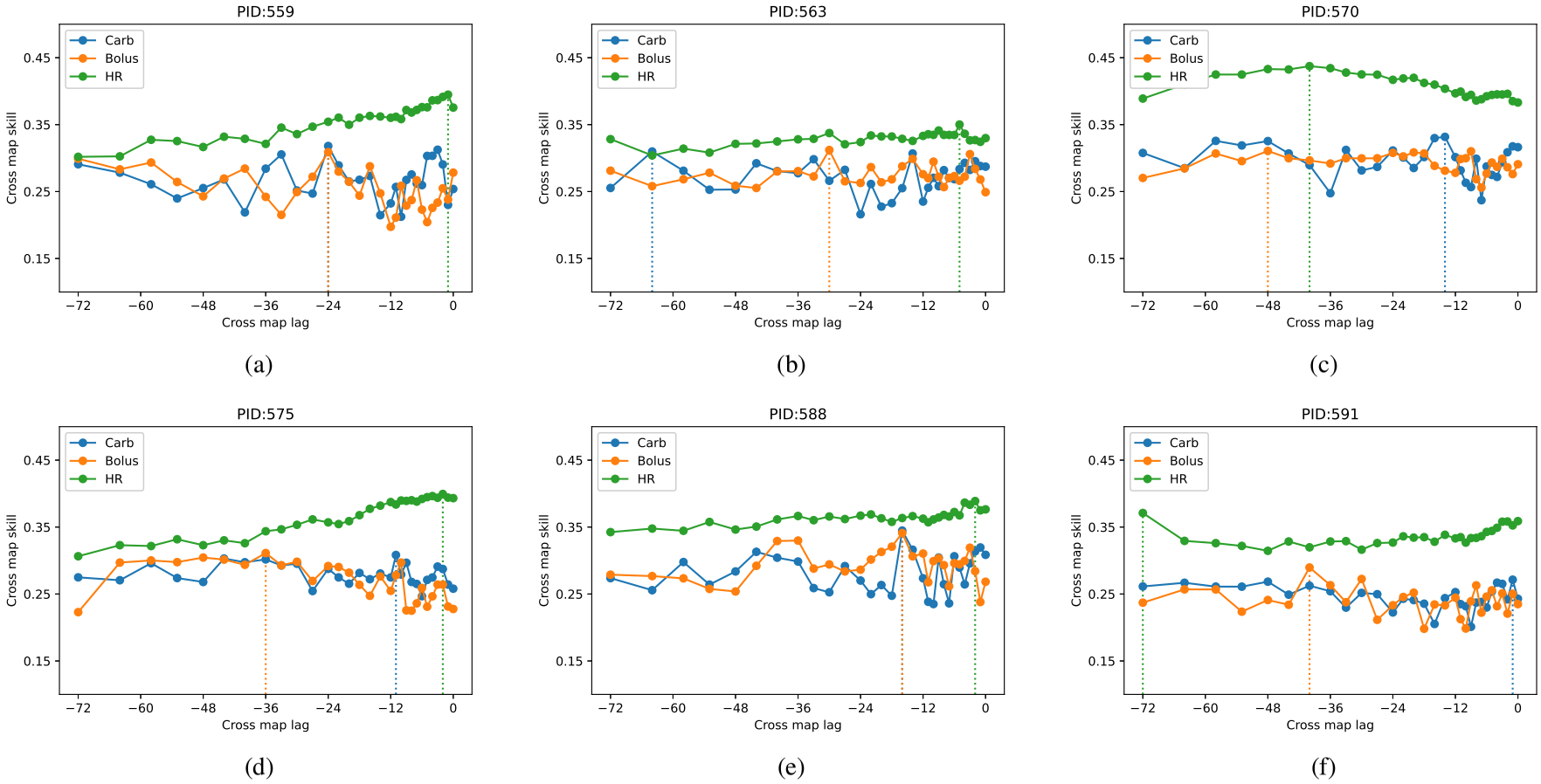


Figure 4.4: The cross map skill as a function of lag for PID (a) 559, (b) 563, (c) 570, (d) 575, (e) 588, and (f) 591 in Ohio_2018 dataset. The points show average values over 10 sampled time series for each lag. The vertical dotted lines show the maximum amount of cross map skill and related time lag.

Note. PID: Patient identity; Carb: Carbohydrate intake, Bolus: Injected bolus insulin; HR: Heart rate.

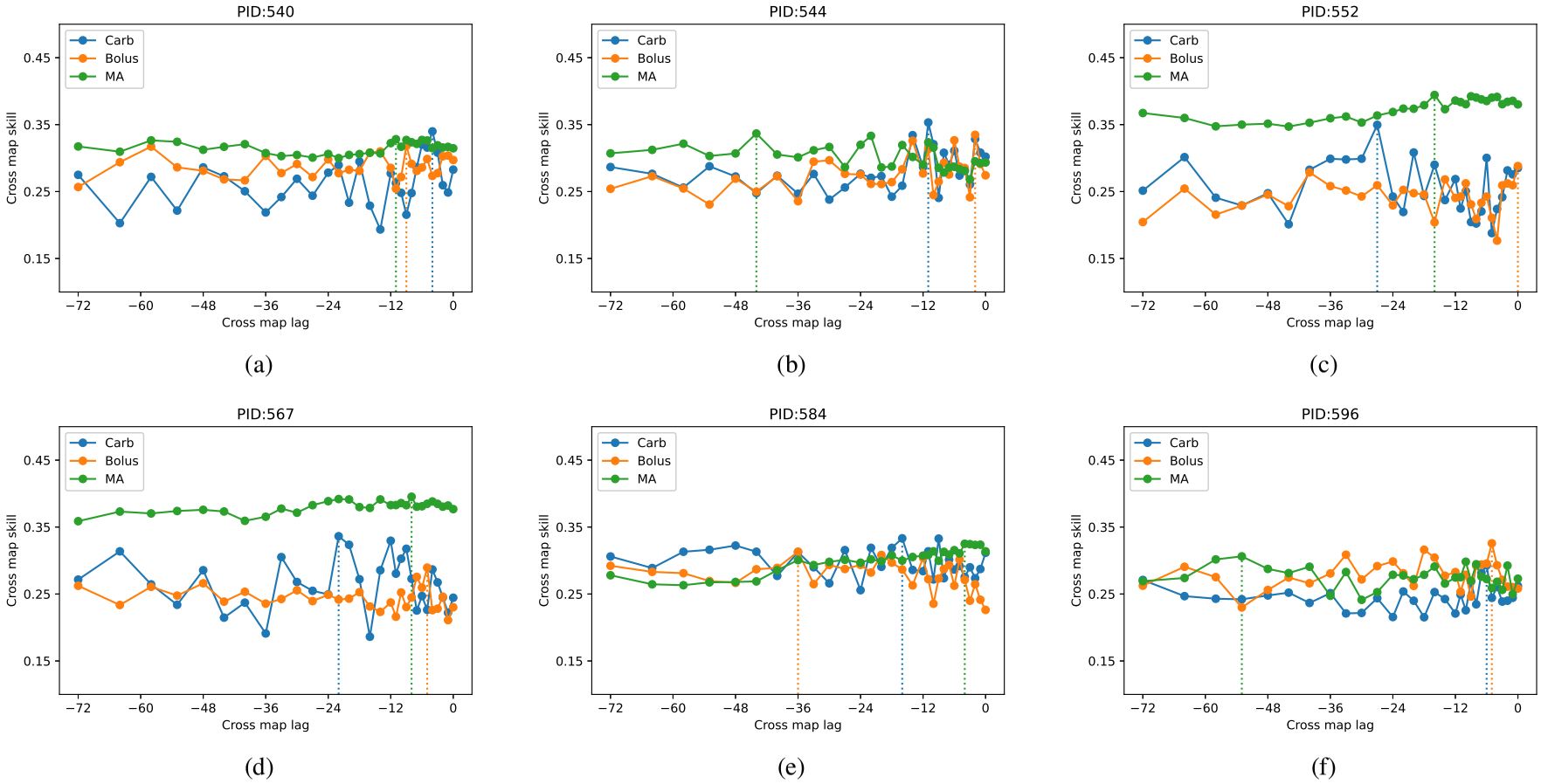


Figure 4.5: The cross map skill as a function of lag for PID (a) 540, (b) 544, (c) 552, (d) 567, (e) 584, and (f) 596 in Ohio_2020 dataset. The points show average values over 10 sampled time series for each lag. The vertical dotted lines show the maximum amount of cross map skill and related time lag.

Note. PID: Patient identity; Carb: Carbohydrate intake, Bolus: Injected bolus insulin; MA: Magnitude of acceleration.

Table 4.3: The results of optimal lag corresponding to the maximum cross map skill values for carbohydrate, bolus, and HR in the Ohio_2018 and Ohio_2020 datasets.

	PID	Carb		Bolus		HR	
		τ_{opt}	ρ_{max}	τ_{opt}	ρ_{max}	τ_{opt}	ρ_{max}
Ohio_2018	559	24	0.318 ± 0.042	24	0.309 ± 0.033	1	0.395 ± 0.022
	563	64	0.310 ± 0.034	30	0.312 ± 0.017	5	0.350 ± 0.008
	570	14	0.332 ± 0.025	48	0.311 ± 0.020	40	0.437 ± 0.013
	575	11	0.308 ± 0.015	36	0.311 ± 0.043	2	0.399 ± 0.009
	588	16	0.345 ± 0.088	16	0.341 ± 0.034	2	0.389 ± 0.007
	591	1	0.271 ± 0.032	40	0.290 ± 0.029	72	0.371 ± 0.021
	PID	Carb		Bolus		MA	
		τ_{opt}	ρ_{max}	τ_{opt}	ρ_{max}	τ_{opt}	ρ_{max}
Ohio_2020	540	4	0.340 ± 0.026	9	0.318 ± 0.028	11	0.328 ± 0.007
	544	11	0.353 ± 0.031	2	0.335 ± 0.026	44	0.337 ± 0.023
	552	27	0.350 ± 0.050	0	0.288 ± 0.029	16	0.394 ± 0.013
	567	22	0.336 ± 0.057	5	0.289 ± 0.015	8	0.395 ± 0.010
	584	16	0.333 ± 0.031	36	0.313 ± 0.021	4	0.325 ± 0.015
	596	6	0.293 ± 0.016	5	0.326 ± 0.033	53	0.306 ± 0.015

Note. PID: Patient identity; Carb: Carbohydrate intake; Bolus: Injected bolus insulin; HR: Heart rate; MA: Magnitude of acceleration; ρ_{max}: Maximum cross map skill; τ_{opt}: Optimal cross map lag.

4.3.2 Leveraging causality in BGL prediction

The results of evaluation criteria and statistical analyses for assessing the performance of BGL prediction with and without leveraging causality are presented in the following. Due to the stochastic nature of the neural networks, with performance depending on random initialisation, the prediction models were run five times for each approach, and the average of the results over the five repetitions is reported.

Tables 4.4 and 4.5 provide the average and standard deviation of evaluation criteria over five runs for the three BGL prediction approaches in Ohio_2018 for prediction horizons of 30 and 60 minutes, respectively. Similarly, Tables 4.6 and 4.7 show the evaluation criteria in Ohio_2020 for prediction horizons of 30 and 60 minutes, respectively.

In each table, the average of evaluation metrics was calculated over data providers for each approach and for each prediction horizon. These values could be used for comparing the performances of different approaches. However, since considering just these averages would be arguable to have a conclusive comparison, the statistical analyses results were also considered [86].

Table 4.4: Evaluation results of the prediction models for different approaches in Ohio_2018 dataset for the prediction horizon of 30 minutes.

PID	Approach	Model	RMSE	MAE	MCC	SE
559	Normal	LSTM	20.78 ± 0.22	14.79 ± 0.22	0.79 ± 0.01	0.20 ± 0.00
		CRNN	20.85 ± 0.03	14.46 ± 0.10	0.79 ± 0.00	0.20 ± 0.00
	CCMBA	LSTM	19.72 ± 0.40	13.89 ± 0.41	0.80 ± 0.01	0.19 ± 0.01
		CRNN	20.50 ± 0.20	14.22 ± 0.13	0.79 ± 0.00	0.20 ± 0.00
	ECCMBA	LSTM	20.19 ± 0.44	14.26 ± 0.31	0.79 ± 0.00	0.20 ± 0.00
		CRNN	21.47 ± 0.26	14.90 ± 0.16	0.77 ± 0.00	0.21 ± 0.00
563	Normal	LSTM	20.74 ± 0.17	14.19 ± 0.18	0.75 ± 0.01	0.19 ± 0.00
		CRNN	19.50 ± 0.05	14.01 ± 0.12	0.74 ± 0.01	0.20 ± 0.00
	CCMBA	LSTM	19.46 ± 0.31	13.30 ± 0.19	0.76 ± 0.01	0.18 ± 0.00
		CRNN	19.02 ± 0.04	13.53 ± 0.06	0.75 ± 0.00	0.19 ± 0.00
	ECCMBA	LSTM	19.10 ± 0.11	13.24 ± 0.07	0.75 ± 0.00	0.19 ± 0.00
		CRNN	19.36 ± 0.06	13.94 ± 0.13	0.73 ± 0.01	0.20 ± 0.00
570	Normal	LSTM	18.05 ± 0.50	12.42 ± 0.41	0.86 ± 0.01	0.12 ± 0.00
		CRNN	20.40 ± 0.55	14.49 ± 0.37	0.84 ± 0.01	0.14 ± 0.00
	CCMBA	LSTM	16.64 ± 0.41	11.54 ± 0.25	0.87 ± 0.00	0.11 ± 0.00
		CRNN	19.84 ± 0.69	14.09 ± 0.52	0.84 ± 0.00	0.14 ± 0.00
	ECCMBA	LSTM	16.91 ± 0.13	11.66 ± 0.11	0.86 ± 0.01	0.11 ± 0.00
		CRNN	21.20 ± 0.36	14.87 ± 0.25	0.83 ± 0.00	0.14 ± 0.00
575	Normal	LSTM	25.71 ± 0.74	16.24 ± 0.20	0.71 ± 0.01	0.24 ± 0.00
		CRNN	27.39 ± 0.24	18.41 ± 0.23	0.68 ± 0.01	0.28 ± 0.00
	CCMBA	LSTM	24.73 ± 0.26	15.46 ± 0.09	0.73 ± 0.01	0.23 ± 0.00
		CRNN	26.51 ± 0.20	18.01 ± 0.27	0.69 ± 0.01	0.27 ± 0.00
	ECCMBA	LSTM	25.32 ± 0.43	16.01 ± 0.45	0.73 ± 0.02	0.24 ± 0.01
		CRNN	27.78 ± 0.29	19.19 ± 0.24	0.67 ± 0.01	0.29 ± 0.00
588	Normal	LSTM	18.82 ± 0.26	13.80 ± 0.24	0.76 ± 0.01	0.18 ± 0.00
		CRNN	22.63 ± 0.16	16.51 ± 0.14	0.70 ± 0.00	0.22 ± 0.00
	CCMBA	LSTM	17.88 ± 0.36	13.15 ± 0.22	0.77 ± 0.01	0.17 ± 0.00
		CRNN	21.88 ± 0.30	15.86 ± 0.21	0.70 ± 0.00	0.21 ± 0.00
	ECCMBA	LSTM	18.80 ± 0.24	13.58 ± 0.17	0.77 ± 0.00	0.18 ± 0.00
		CRNN	23.28 ± 0.28	16.68 ± 0.21	0.69 ± 0.00	0.22 ± 0.00

(continued on next page)

Table 4.4 (continued)

PID	Approach	Model	RMSE	MAE	MCC	SE
591	Normal	LSTM	23.05 ± 0.80	16.75 ± 0.78	0.64 ± 0.02	0.28 ± 0.01
		CRNN	26.95 ± 1.22	20.14 ± 1.07	0.56 ± 0.03	0.33 ± 0.02
	CCMBA	LSTM	22.13 ± 0.43	16.02 ± 0.32	0.64 ± 0.01	0.28 ± 0.01
		CRNN	24.20 ± 0.18	17.84 ± 0.14	0.63 ± 0.01	0.29 ± 0.00
	ECCMBA	LSTM	22.37 ± 0.21	16.31 ± 0.25	0.64 ± 0.01	0.28 ± 0.00
		CRNN	24.64 ± 0.71	18.21 ± 0.63	0.61 ± 0.03	0.30 ± 0.01
Avg	Normal	LSTM	21.19 ± 0.45	14.70 ± 0.34	0.75 ± 0.01	0.20 ± 0.00
		CRNN	22.95 ± 0.38	16.34 ± 0.34	0.72 ± 0.01	0.23 ± 0.00
	CCMBA	LSTM	20.09 ± 0.36	13.90 ± 0.25	0.76 ± 0.01	0.19 ± 0.00
		CRNN	21.99 ± 0.27	15.59 ± 0.22	0.73 ± 0.00	0.22 ± 0.00
	ECCMBA	LSTM	20.45 ± 0.26	14.18 ± 0.23	0.76 ± 0.01	0.20 ± 0.00
		CRNN	22.96 ± 0.33	16.30 ± 0.27	0.72 ± 0.01	0.23 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CCMBA: Convergent cross mapping based approach; ECCMBA: Extended convergent cross mapping based approach.

Table 4.5: Evaluation results of the prediction models for different approaches in Ohio_2018 dataset for the prediction horizon of 60 minutes.

PID	Approach	Model	RMSE	MAE	MCC	SE
559	Normal	LSTM	33.73 ± 0.38	24.54 ± 0.27	0.64 ± 0.01	0.33 ± 0.00
		CRNN	31.97 ± 0.40	22.92 ± 0.41	0.66 ± 0.01	0.31 ± 0.00
	CCMBA	LSTM	32.59 ± 0.81	23.75 ± 0.53	0.64 ± 0.01	0.32 ± 0.01
		CRNN	31.52 ± 0.16	22.55 ± 0.15	0.65 ± 0.01	0.31 ± 0.00
	ECCMBA	LSTM	34.97 ± 0.47	26.12 ± 0.49	0.62 ± 0.01	0.35 ± 0.00
		CRNN	35.73 ± 0.10	25.69 ± 0.09	0.59 ± 0.00	0.35 ± 0.00
563	Normal	LSTM	33.77 ± 0.45	23.97 ± 0.40	0.54 ± 0.01	0.32 ± 0.01
		CRNN	31.08 ± 0.09	22.12 ± 0.16	0.58 ± 0.01	0.30 ± 0.00
	CCMBA	LSTM	32.44 ± 0.47	23.40 ± 0.53	0.52 ± 0.03	0.32 ± 0.01
		CRNN	30.73 ± 0.08	21.86 ± 0.10	0.58 ± 0.01	0.30 ± 0.00
	ECCMBA	LSTM	31.34 ± 0.37	22.47 ± 0.25	0.54 ± 0.01	0.31 ± 0.00
		CRNN	31.10 ± 0.21	22.34 ± 0.12	0.54 ± 0.01	0.31 ± 0.00

(continued on next page)

Table 4.5 (continued)

PID	Approach	Model	RMSE	MAE	MCC	SE
570	Normal	LSTM	29.92 ± 0.48	21.64 ± 0.39	0.79 ± 0.01	0.20 ± 0.00
		CRNN	29.79 ± 0.08	21.81 ± 0.04	0.77 ± 0.00	0.21 ± 0.00
	CCMBA	LSTM	28.87 ± 0.56	20.78 ± 0.49	0.80 ± 0.01	0.19 ± 0.00
		CRNN	30.22 ± 0.30	22.05 ± 0.19	0.76 ± 0.00	0.21 ± 0.00
	ECCMBA	LSTM	29.38 ± 0.35	20.93 ± 0.15	0.79 ± 0.01	0.20 ± 0.01
		CRNN	31.13 ± 0.33	22.58 ± 0.18	0.76 ± 0.00	0.22 ± 0.00
575	Normal	LSTM	39.06 ± 0.77	27.33 ± 0.55	0.51 ± 0.02	0.41 ± 0.01
		CRNN	37.65 ± 0.05	26.47 ± 0.08	0.53 ± 0.01	0.39 ± 0.00
	CCMBA	LSTM	37.56 ± 0.24	26.48 ± 0.20	0.53 ± 0.01	0.39 ± 0.00
		CRNN	37.66 ± 0.43	26.57 ± 0.54	0.52 ± 0.01	0.39 ± 0.01
	ECCMBA	LSTM	38.53 ± 0.53	27.39 ± 0.57	0.52 ± 0.01	0.41 ± 0.01
		CRNN	38.48 ± 0.22	27.57 ± 0.27	0.50 ± 0.01	0.41 ± 0.01
588	Normal	LSTM	31.55 ± 0.34	22.94 ± 0.24	0.58 ± 0.01	0.29 ± 0.00
		CRNN	33.01 ± 0.10	24.34 ± 0.04	0.54 ± 0.01	0.32 ± 0.00
	CCMBA	LSTM	30.83 ± 0.49	22.61 ± 0.35	0.59 ± 0.01	0.29 ± 0.00
		CRNN	33.50 ± 0.98	24.65 ± 0.68	0.55 ± 0.01	0.32 ± 0.01
	ECCMBA	LSTM	32.18 ± 0.38	23.30 ± 0.30	0.58 ± 0.00	0.30 ± 0.00
		CRNN	35.19 ± 0.18	25.49 ± 0.15	0.51 ± 0.00	0.33 ± 0.00
591	Normal	LSTM	35.97 ± 0.83	27.79 ± 0.63	0.43 ± 0.02	0.44 ± 0.01
		CRNN	42.00 ± 1.65	32.98 ± 1.81	0.28 ± 0.05	0.53 ± 0.03
	CCMBA	LSTM	35.11 ± 0.70	27.50 ± 0.54	0.44 ± 0.02	0.43 ± 0.01
		CRNN	39.12 ± 0.24	30.13 ± 0.23	0.36 ± 0.01	0.48 ± 0.00
	ECCMBA	LSTM	34.79 ± 0.14	27.21 ± 0.23	0.43 ± 0.02	0.43 ± 0.00
		CRNN	39.36 ± 0.19	30.39 ± 0.20	0.35 ± 0.00	0.48 ± 0.00
Avg	Normal	LSTM	34.00 ± 0.54	24.70 ± 0.41	0.58 ± 0.02	0.33 ± 0.01
		CRNN	34.25 ± 0.40	25.11 ± 0.42	0.56 ± 0.02	0.34 ± 0.01
	CCMBA	LSTM	32.90 ± 0.54	24.09 ± 0.44	0.59 ± 0.01	0.32 ± 0.01
		CRNN	33.79 ± 0.37	24.63 ± 0.32	0.57 ± 0.01	0.34 ± 0.00
	ECCMBA	LSTM	33.53 ± 0.37	24.57 ± 0.33	0.58 ± 0.01	0.33 ± 0.00

		CRNN	35.16 ± 0.21	25.68 ± 0.17	0.54 ± 0.00	0.35 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CCMBA: Convergent cross mapping based approach; ECCMBA: Extended convergent cross mapping based approach.

Table 4.6: Evaluation results of the prediction models for different approaches in Ohio_2020 dataset for the prediction horizon of 30 minutes.

PID	Approach	Model	RMSE	MAE	MCC	SE
540	Normal	LSTM	21.68 ± 0.51	16.17 ± 0.21	0.70 ± 0.01	0.26 ± 0.00
		CRNN	27.83 ± 0.50	20.71 ± 0.36	0.65 ± 0.02	0.31 ± 0.01
	CCMBA	LSTM	21.33 ± 0.55	15.98 ± 0.37	0.71 ± 0.00	0.26 ± 0.00
		CRNN	26.60 ± 0.45	19.70 ± 0.35	0.68 ± 0.01	0.29 ± 0.01
	ECCMBA	LSTM	21.29 ± 0.12	16.02 ± 0.09	0.70 ± 0.00	0.25 ± 0.00
		CRNN	27.46 ± 0.46	20.22 ± 0.35	0.67 ± 0.01	0.30 ± 0.01
544	Normal	LSTM	19.50 ± 0.21	13.49 ± 0.23	0.78 ± 0.00	0.20 ± 0.00
		CRNN	21.33 ± 0.14	15.15 ± 0.11	0.74 ± 0.00	0.21 ± 0.00
	CCMBA	LSTM	18.51 ± 0.72	12.93 ± 0.39	0.79 ± 0.02	0.19 ± 0.01
		CRNN	21.10 ± 0.39	14.93 ± 0.30	0.75 ± 0.00	0.21 ± 0.01
	ECCMBA	LSTM	20.05 ± 0.96	13.49 ± 0.22	0.80 ± 0.01	0.20 ± 0.00
		CRNN	21.86 ± 0.51	15.12 ± 0.39	0.76 ± 0.00	0.21 ± 0.00
552	Normal	LSTM	18.42 ± 0.49	13.56 ± 0.22	0.70 ± 0.01	0.23 ± 0.00
		CRNN	20.33 ± 0.46	14.48 ± 0.29	0.69 ± 0.01	0.24 ± 0.00
	CCMBA	LSTM	16.77 ± 0.10	12.44 ± 0.07	0.73 ± 0.00	0.21 ± 0.00
		CRNN	19.26 ± 0.42	13.68 ± 0.35	0.73 ± 0.02	0.22 ± 0.01
	ECCMBA	LSTM	16.84 ± 0.23	12.68 ± 0.27	0.74 ± 0.00	0.21 ± 0.00
		CRNN	20.61 ± 1.07	14.64 ± 0.74	0.70 ± 0.03	0.24 ± 0.01
567	Normal	LSTM	20.85 ± 0.14	14.73 ± 0.17	0.64 ± 0.02	0.25 ± 0.01
		CRNN	24.80 ± 0.08	17.62 ± 0.07	0.64 ± 0.02	0.28 ± 0.00
	CCMBA	LSTM	20.66 ± 0.16	14.64 ± 0.26	0.65 ± 0.02	0.25 ± 0.01
		CRNN	24.46 ± 0.44	17.29 ± 0.32	0.66 ± 0.01	0.27 ± 0.00
	ECCMBA	LSTM	20.70 ± 0.15	14.74 ± 0.11	0.65 ± 0.01	0.25 ± 0.00
		CRNN	26.78 ± 0.43	18.96 ± 0.28	0.63 ± 0.02	0.29 ± 0.01
584	Normal	LSTM	21.96 ± 0.18	16.03 ± 0.12	0.77 ± 0.00	0.22 ± 0.00

(continued on next page)

Table 4.6 (continued)

PID	Approach	Model	RMSE	MAE	MCC	SE
		CRNN	24.28 ± 0.04	17.57 ± 0.04	0.73 ± 0.00	0.24 ± 0.00
	CCMBA	LSTM	21.73 ± 0.44	15.89 ± 0.47	0.77 ± 0.00	0.22 ± 0.01
		CRNN	23.65 ± 0.33	17.12 ± 0.28	0.74 ± 0.01	0.23 ± 0.00
	ECCMBA	LSTM	23.12 ± 0.27	17.01 ± 0.24	0.75 ± 0.00	0.23 ± 0.00
		CRNN	25.12 ± 0.35	18.18 ± 0.24	0.72 ± 0.00	0.25 ± 0.00
596	Normal	LSTM	17.97 ± 0.31	12.77 ± 0.31	0.76 ± 0.01	0.20 ± 0.00
		CRNN	19.80 ± 0.21	14.24 ± 0.18	0.73 ± 0.00	0.22 ± 0.00
	CCMBA	LSTM	17.63 ± 0.11	12.59 ± 0.09	0.76 ± 0.00	0.20 ± 0.00
		CRNN	18.81 ± 0.04	13.53 ± 0.03	0.73 ± 0.01	0.21 ± 0.00
	ECCMBA	LSTM	18.00 ± 0.09	12.85 ± 0.12	0.76 ± 0.01	0.20 ± 0.00
		CRNN	19.94 ± 0.27	14.32 ± 0.22	0.72 ± 0.00	0.23 ± 0.00
Avg	Normal	LSTM	20.06 ± 0.31	14.46 ± 0.21	0.73 ± 0.01	0.23 ± 0.00
		CRNN	23.06 ± 0.24	16.63 ± 0.17	0.70 ± 0.01	0.25 ± 0.00
	CCMBA	LSTM	19.44 ± 0.35	14.08 ± 0.27	0.74 ± 0.01	0.22 ± 0.00
		CRNN	22.31 ± 0.35	16.04 ± 0.27	0.72 ± 0.01	0.24 ± 0.00
	ECCMBA	LSTM	20.00 ± 0.31	14.47 ± 0.18	0.73 ± 0.01	0.23 ± 0.00
		CRNN	23.63 ± 0.52	16.91 ± 0.37	0.70 ± 0.01	0.25 ± 0.01

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CCMBA: Convergent cross mapping based approach; ECCMBA: Extended convergent cross mapping based approach.

Table 4.7: Evaluation results of the prediction models for different approaches in Ohio_2020 dataset for the prediction horizon of 60 minutes.

PID	Approach	Model	RMSE	MAE	MCC	SE
540	Normal	LSTM	41.88 ± 0.94	31.10 ± 0.44	0.52 ± 0.02	0.45 ± 0.00
		CRNN	42.93 ± 0.16	32.28 ± 0.10	0.52 ± 0.01	0.46 ± 0.00
	CCMBA	LSTM	40.90 ± 0.51	30.95 ± 0.25	0.52 ± 0.01	0.44 ± 0.00
		CRNN	42.05 ± 0.14	31.64 ± 0.10	0.53 ± 0.01	0.45 ± 0.00
	ECCMBA	LSTM	40.82 ± 0.35	31.28 ± 0.21	0.52 ± 0.01	0.45 ± 0.01
		CRNN	44.56 ± 0.20	33.40 ± 0.15	0.48 ± 0.01	0.48 ± 0.00

(continued on next page)

Table 4.7 (continued)

PID	Approach	Model	RMSE	MAE	MCC	SE
544	Normal	LSTM	32.65 ± 0.42	23.13 ± 0.36	0.64 ± 0.02	0.32 ± 0.01
		CRNN	33.82 ± 0.18	24.33 ± 0.19	0.54 ± 0.01	0.34 ± 0.00
	CCMBA	LSTM	31.92 ± 0.59	22.70 ± 0.30	0.63 ± 0.01	0.32 ± 0.00
		CRNN	32.38 ± 0.17	23.25 ± 0.13	0.56 ± 0.01	0.32 ± 0.00
	ECCMBA	LSTM	34.44 ± 0.78	24.34 ± 0.30	0.63 ± 0.01	0.34 ± 0.00
		CRNN	34.73 ± 0.53	24.45 ± 0.36	0.56 ± 0.01	0.34 ± 0.00
552	Normal	LSTM	31.71 ± 0.78	23.82 ± 0.71	0.56 ± 0.02	0.38 ± 0.01
		CRNN	32.50 ± 0.07	22.77 ± 0.05	0.54 ± 0.01	0.36 ± 0.00
	CCMBA	LSTM	30.05 ± 0.16	22.60 ± 0.13	0.58 ± 0.01	0.36 ± 0.00
		CRNN	32.49 ± 0.43	22.77 ± 0.21	0.55 ± 0.00	0.36 ± 0.00
	ECCMBA	LSTM	30.19 ± 0.21	22.70 ± 0.25	0.57 ± 0.02	0.36 ± 0.00
		CRNN	34.72 ± 1.60	24.84 ± 1.56	0.54 ± 0.02	0.39 ± 0.02
567	Normal	LSTM	37.43 ± 0.32	27.64 ± 0.15	0.37 ± 0.01	0.47 ± 0.00
		CRNN	43.38 ± 0.49	31.38 ± 0.57	0.34 ± 0.02	0.49 ± 0.01
	CCMBA	LSTM	36.88 ± 0.31	27.75 ± 0.39	0.37 ± 0.01	0.47 ± 0.01
		CRNN	41.27 ± 0.44	29.87 ± 0.26	0.36 ± 0.02	0.46 ± 0.00
	ECCMBA	LSTM	37.18 ± 0.49	28.14 ± 0.44	0.37 ± 0.01	0.47 ± 0.01
		CRNN	41.94 ± 0.34	30.22 ± 0.22	0.37 ± 0.01	0.47 ± 0.00
584	Normal	LSTM	39.99 ± 0.95	31.04 ± 0.79	0.53 ± 0.02	0.42 ± 0.01
		CRNN	37.91 ± 0.07	27.58 ± 0.10	0.56 ± 0.00	0.36 ± 0.00
	CCMBA	LSTM	40.81 ± 0.56	31.49 ± 0.43	0.53 ± 0.01	0.42 ± 0.01
		CRNN	37.58 ± 0.38	27.31 ± 0.31	0.57 ± 0.01	0.36 ± 0.00
	ECCMBA	LSTM	40.03 ± 1.84	30.41 ± 1.61	0.57 ± 0.03	0.40 ± 0.02
		CRNN	40.54 ± 0.27	29.07 ± 0.21	0.55 ± 0.00	0.38 ± 0.00
596	Normal	LSTM	29.83 ± 1.24	22.17 ± 0.89	0.60 ± 0.02	0.32 ± 0.01
		CRNN	31.77 ± 0.17	23.17 ± 0.16	0.57 ± 0.00	0.34 ± 0.00
	CCMBA	LSTM	28.18 ± 0.45	20.89 ± 0.45	0.62 ± 0.01	0.31 ± 0.01
		CRNN	30.09 ± 1.19	21.87 ± 0.91	0.59 ± 0.02	0.32 ± 0.01
	ECCMBA	LSTM	29.44 ± 0.13	21.68 ± 0.17	0.59 ± 0.00	0.32 ± 0.00
		CRNN	32.05 ± 0.43	23.29 ± 0.29	0.54 ± 0.01	0.34 ± 0.00
Avg	Normal	LSTM	35.58 ± 0.77	26.48 ± 0.56	0.53 ± 0.02	0.39 ± 0.01

(continued on next page)

Table 4.7 (continued)

PID	Approach	Model	RMSE	MAE	MCC	SE
		CRNN	37.05 ± 0.19	26.92 ± 0.20	0.51 ± 0.01	0.39 ± 0.00
	CCMBA	LSTM	34.79 ± 0.43	26.06 ± 0.32	0.54 ± 0.01	0.39 ± 0.00
		CRNN	35.98 ± 0.46	26.12 ± 0.32	0.52 ± 0.01	0.38 ± 0.00
	ECCMBA	LSTM	35.35 ± 0.63	26.42 ± 0.49	0.54 ± 0.01	0.39 ± 0.01
		CRNN	38.09 ± 0.56	27.54 ± 0.47	0.51 ± 0.01	0.40 ± 0.01

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CCMBA: Convergent cross mapping based approach; ECCMBA: Extended convergent cross mapping based approach.

The p-values of the Wilcoxon test are shown in Table 4.8. In the test, the null hypothesis was that each of the proposed CCMBA and ECCMBA approaches and the Normal approach had the same distribution. Each LSTM and CRNN model was tested separately as the prediction model. The test was applied to all evaluation metrics, including RMSE, MAE, MCC, and SE for both prediction horizons of 30 and 60 minutes and separately for Ohio_2018 and Ohio_2020 datasets. The significant p-values, based on the significant level of five percent, are marked with bold font.

Considering the results of Tables 4.4 and 4.8, in the Ohio_2018 dataset, and in terms of the LSTM prediction model, it can be concluded that for the prediction horizon of 30 minutes, the CCMBA statistically significantly improved the average evaluation metrics of RMSE, MAE, MCC, and SE over all patients by 5.191%, 5.484%, 1.594%, and 4.433%, respectively, compared to the Normal approach. In addition, the ECCMBA improved the performance of BGL prediction for the average evaluation metrics of RMSE, MAE, and SE by 3.506%, 3.545%, and 1.970%, respectively, compared to the Normal approach. Also, when assigning CRNN as the prediction model, the CCMBA statistically significantly outperformed the Normal approach for the prediction horizon of 30 minutes for the average evaluation metrics of RMSE, MAE, and SE over all patients by 4.187%, 4.554%, and 4.846%, respectively. Similarly, considering Tables 4.5 and 4.8, it can be inferred that for the prediction horizon of 60 minutes, by assigning LSTM as the prediction model, the CCMBA compared to the Normal approach, improved RMSE, MAE, and SE metrics by 3.238%, 2.490%, and 2.115%, respectively.

The results in Tables 4.6 and 4.8 show that in the Ohio_2020 dataset and for the prediction horizon of 30 minutes, in the case of LSTM as the prediction model, it

Table 4.8: P-values of the Wilcoxon test for comparing the evaluation metrics of BGL prediction models using CCMBA and ECCMBA with Normal approach 30 and 60 minutes in advance over the individuals in Ohio_2018 and Ohio_2020 datasets.

	PH	Approach	Model	RMSE	MAE	MCC	SE
Ohio_2018	30 min	CCMBA	LSTM	**0.031**	**0.031**	**0.031**	**0.031**
			CRNN	**0.031**	**0.031**	0.094	**0.031**
		ECCMBA	LSTM	**0.031**	**0.031**	0.062	**0.031**
			CRNN	0.562	0.562	0.438	0.438
	60 min	CCMBA	LSTM	**0.031**	**0.031**	0.438	**0.031**
			CRNN	0.844	0.562	0 1.000	0.844
		ECCMBA	LSTM	0.688	0.844	1.000	1.000
			CRNN	0.312	0.312	0.438	0.438
Ohio_2020	30 min	CCMBA	LSTM	**0.031**	**0.031**	0.156	**0.031**
			CRNN	**0.031**	**0.031**	**0.031**	**0.031**
		ECCMBA	LSTM	1.000	0.844	0.438	0.844
			CRNN	0.156	0.312	0.844	0.438
	60 min	CCMBA	LSTM	0.156	0.312	0.438	**0.031**
			CRNN	**0.031**	**0.031**	**0.031**	**0.031**
		ECCMBA	LSTM	0.562	1.000	0.438	0.688
			CRNN	0.156	0.219	0.844	0.438

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CCMBA: Convergent cross mapping based approach; ECCMBA: Extended convergent cross mapping based approach.

can be inferred that the CCMBA statistically significantly improved RMSE, MAE, and SE metrics, for the prediction horizon of 30 minutes compared to the Normal BGL prediction approach with the improvement percentage of 3.101%, 2.621%, and 2.203%, respectively. Also, with CRNN as the prediction model, the CCMBA statistically significantly outperformed the Normal BGL prediction approach in terms of all RMSE, MAE, MCC, and SE metrics, with improvement percentages of 3.252%, 3.536%, 2.726%, and 4.000%. Also, the results in Tables 4.7 and 4.8 show that for the prediction horizon of 60 minutes, the CCMBA improved the SE metric by 1.777% compared to the Normal BGL prediction approach, in the case of LSTM as the prediction model. Moreover, with CRNN as the prediction model, the CCMBA outperformed the Normal approach in terms of all RMSE, MAE, MCC, and SE metrics, with improvement percentages of 2.901%, 2.972%, 2.539%, and 3.061% , respectively.

Overall, the results indicate that causality inference can be useful in improving

BGL prediction performance and, more specifically, highlight the superior performance of the CCMBA compared to the Normal approach. The ECCMBA, however, was not very effective.

4.4 Summary

The work presented in this chapter investigated relations between BGL and affecting variables of T1DM management, including carbohydrate, bolus, and PA (HR and MA) in a causality framework. CCM was applied to quantify the causality strength for each variable. ECCM was used to examine the optimal time lag for the impact of each variable by investigating the causality strengths of variables for different lags. Also, in this work, the feasibility of leveraging causality information as prior knowledge for BGL prediction was investigated. To do so, two BGL prediction models (LSTM and CRNN) were applied utilising two publicly accessible Ohio datasets to forecast BGL 30 and 60 minutes in advance. Then, two causality-based approaches (CCMBA and ECCMBA) were proposed for integrating the causality information in BGL prediction. In the CCMBA, the causality strengths were used as weights for variables. In the ECCMBA, the optimal lag of causation was used as a shift for each variable weighted by the corresponding causal strength.

Applying CCM analysis illustrates activity data has stronger causality with BGL data compared to carbohydrate or bolus data. This supports the effectiveness of integrating activity data with CGM data to improve T1DM management systems, such as systems providing predictive alarms or closed-loop systems where BGL prediction is used to automatically alter the rate of insulin infusion in insulin pumps. Moreover, the results indicate that CCMBA is a more effective approach than ECCMBA for deploying causality information in BGL prediction.

Chapter 5

Leveraging physical activity in blood glucose level prediction

5.1 Preface

As previously described, PA plays an important role in managing T1DM. Due to insufficient explicit knowledge of how to translate the impact of PA on BGL, optimal management is challenging and it is difficult for clinicians to provide patients with specific advice concerning PA [12]. Part of the complexity arises because it has recently been shown that BGL can vary significantly for individuals during and after exercise from one day to another, even for the same type and duration of the exercise performed at the same time of day and after consuming similar meals [30]. Hence, although regular exercise is beneficial for T1DM patients as it helps to reduce the risk of cardiovascular disease, maintaining normoglycaemia is challenging. Indeed, many people with T1DM avoid exercise so as to not increase the chances of hyperglycaemia or hypoglycaemia events before or after exercise [11, 159, 160]. Although limited works have been performed considering PA in BGL prediction, there is still a demand to discover optimal approaches for incorporating PA in BGL prediction. Accordingly, it is beneficial to perform a rigorous investigation by finding information from PA data and examining efficient ways to combine this with BGL in order to improve the performance of BGL prediction.

The work presented in this chapter proposes different approaches for extracting PA information and examines the effectiveness of various levels of PA and BGL data fusion, including signal-level fusion, feature-level fusion, and decision-level fusion. For each fusion level, different approaches are developed to extract information related to PA. In order to have a valid and conclusive deduction, rigorous statistical analyses and evaluation are performed to compare PA-fusion approaches with each

other and with the no-fusion approach to find effective PA fusion approaches.

5.2 Material and methods

5.2.1 Dataset

The dataset used in this work was the Ohio_2018 [66]. As described in Section 2.8, the dataset contained data from physiological sensors and self-reported life events of six individuals with T1DM. In this dataset, BGL data was collected with 5-minute aggregation using a Medtronic Enlite CGM sensor. PA data was automatically recorded as heart rate (HR), step count (SC), galvanic skin response (GSR), and skin temperature (ST) using a Basis Peak band with 5-minute aggregation. Also, patients reported times and duration of sleep, work, and exercise. Individuals' subjective assessments of physical effort for work and exercise on a scale from one to 10, with 10 indicating the highest level of PA, were also used in this work.

In this work, CGM data along with automatic-recorded and self-reported data related to PA were used. The number of data points for BGL and automatic-recorded PA data has been provided in Table 2.3. Also, Table 5.1 shows the count of self-reported data, related to PA with patients' subjective assessment of intensity level. Figure 5.1 shows BGL and PA-related data for a duration of 24 hours of training data for one of the data contributors.

Table 5.1: The number and patients' subjective assessment of intensity levels of PA data.

PID	No data	Sleep	Work/Exercise									
			1	2	3	4	5	6	7	8	9	10
559	7406	4771	0	57	227	810	1142	336	0	0	0	0
563	8107	3021	0	143	2737	462	249	144	0	0	0	0
570	5765	4582	718	2037	285	153	644	254	0	0	0	0
575	7554	4699	0	112	613	1386	1052	443	10	0	0	0
588	5642	5403	0	0	0	1148	3215	404	0	0	0	0
591	10983	4390	0	0	53	53	78	43	34	13	5	0

5.2.2 Preprocessing

First, the missing data had to be dealt with in the preprocessing phase. To do so, the missing BGL, HR, GSR and ST data were imputed using a linear approach where interpolation and extrapolation techniques were used in the training and testing sets,

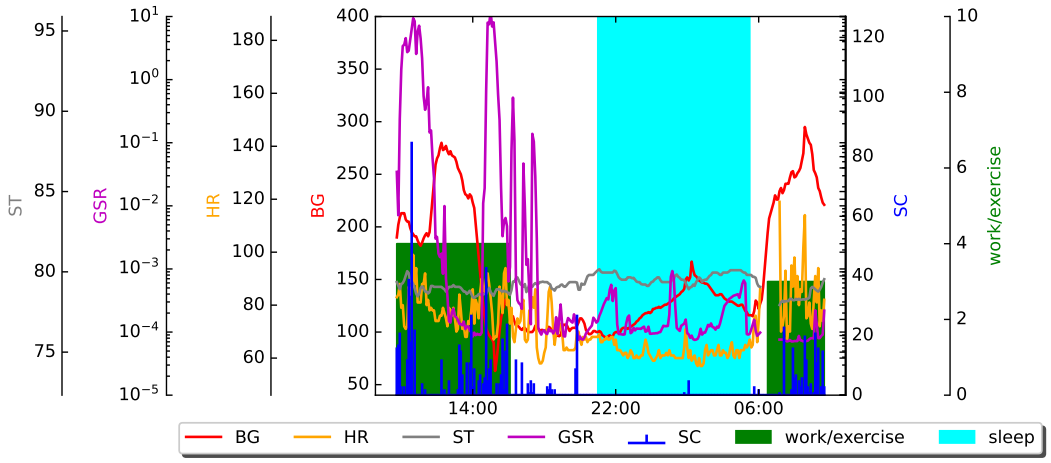


Figure 5.1: BGL and PA-related data for PID 559.

respectively. Also, missing SC data were filled with zero values for non-reported data timestamps. Aligning BGL and PA data was the next preprocessing step. Additionally, since CGM sensors and activity bands were worn at different times, there was a number of data unavailable at the beginning or end of each set. In order to have more reliable data, these timestamps were excluded from the analysis. Moreover, the time series forecasting task of BGL prediction was recast to a supervised learning task. To do so, the time series data were sampled, with lag observations serving as input and future observations serving as output. Then, using a 5-minute rolling window, history samples with 60-minute lengths were assigned as input and future samples with 30-minute and 60-minute lengths were assigned as output. Finally, the scaling process was applied to the input samples for each variable based on its minimum and maximum values over the entire training set.

5.2.3 BGL prediction model

LSTM networks, a type of recurrent neural network, are capable of predicting BGL [83, 95]. The LSTM model developed and described in Sections 3.2.3.2 and 4.2.4.1, was used for the BGL prediction task in the current work. There were three layers in the used vanilla LSTM network: a 200-unit LSTM layer, a 100-unit dense layer, and an output layer of one unit dense layer. The initialiser, activation function, optimiser, and loss function were chosen as He uniform, ReLU, Adam and MSE, respectively. Also, the epoch size was set at 200 and the batch size at 32. The univariate LSTM model was used for using only BGL for prediction, called the no-fusion approach. Furthermore, both univariate and multivariate LSTM models were used for PA-informed approaches, according to the fusion approach. More details

about the model architecture and optimization for univariate and multivariate can be found in [71] and [87], respectively.

5.2.4 Data fusion of PA and BGL

From a data fusion perspective, fusing approaches are categorised into three levels: signal-level, feature-level, and decision-level [161, 162]. In this work, different kinds of information from PA at different levels were fused with BGL data. The performance of the BGL prediction for different fusion information/levels was investigated and compared with the prediction model without PA information. In the following, the approaches for each level of data fusion are described.

5.2.4.1 Signal-level PA fusion

A signal-level data fusion, which used raw sensor data as inputs, was the lowest level of data fusion. In this approach, the history data of BGL from CGM sensors and the corresponding history of automatically recorded PA data from wristbands were used as input for the multivariate LSTM prediction model in three different combinations of BG+HR (BG and HR data), BG+HRSC (BG, HR, and SC data), and BG+Band (BG, HR, SC, GSR, and ST data).

5.2.4.2 Feature-level PA fusion

In this level of data fusion, features from PA data were extracted and fused with the BGL data. To comprehensively investigate this fusion level, three kinds of feature engineering were utilised: subjective PA categories, objective PA clusters, and statistics of PA data. In the following, these features are briefly described.

Subjective PA categorisation Considering that PA is defined as any motion generated by skeletal muscle that increases energy expenditure, it can be categorised as sedentary, light, moderate, and vigorous in terms of relative effort and expenditure of energy [163]. In the first feature extraction technique, self-reported data related to PA were deployed as PA features. To do so, subjective assessments of participants for physical exertion, which were scaled from one to 10, were categorised into three different intensity levels. In detail, data reported with scales of one, two, and three were assigned to the light category; data reported with scales of four and five were allocated to the moderate category; and data reported with a scale of six or more were categorised as vigorous. Also, non-reported timestamps were assumed as not being active and were assigned to the sedentary category. Moreover, sleep

data was assigned to a separate category. Hence, in total, five categories related to different levels of PA intensities including sleep, sedentary, light, moderate, and vigorous were used as subjective PA features. These features were then employed as input along with the BGL data for the multivariate LSTM prediction model. In short, this approach is called BG+SPA.

Objective PA clustering Another feature engineering approach used to extract PA information to be fused with BGL data was clusters generated by K-means, a commonly used clustering approach in the field of unsupervised learning. The number of clusters considered was five, the same as the number of subjective PA groups described previously. This was an objective feature generation using automatic-recorded PA data collected from the wristbands. Similar to the subjective PA categories, inputs for the multivariate LSTM prediction model included the five different PA clusters along with the BGL data. This approach is also referred to as BG+OPA.

Statistics of PA data In this feature category, the statistics of the automatic-recorded PA data, which have been shown to be effective in the BGL prediction in the literature [117, 164], were used for the fusion with BGL data. To do so, PA statistics including the mean and standard deviation were calculated for all the automatic-recorded PA data and added to the corresponding history of BGL data. This was fed as the input of the univariate LSTM model. It is called the BG+StPA approach.

5.2.4.3 Decision-level PA fusion

The highest level of data fusion is decision fusion, which combines information that has already generated some decisions for a given task. In order to examine this level of data fusion, a method employing stacked ensemble learning [114] was developed based on the idea we proposed in our conference paper [117]. The stacked regression consists of multiple models serving as base-learners and a meta-learner fed by the outputs of the base-learners. In this work, instead of using different models as base-learners, the univariate LSTM model was trained twice, once using BGL data, and once using PA data. Thus, at the first level of learning, primary decisions were generated using BGL and PA data, separately. The decisions of the first layer were then stacked and used as input for the meta-learner, which was chosen as a linear regression model to provide the final prediction. Accordingly, deploying the concept of ensemble learning, PA information was fused with the BGL in a decision-level approach.

Similar to the signal-level data fusion, the three combinations of PA data were chosen to be used for training the base-learner. BG&HR, BG&HRSC, and BG&Band are the names of fusion approaches for fusing BGL with (HR), (HR and SC), and (HR, SC, GSR, and ST), respectively.

5.2.5 Evaluation criteria

In this study, the performance of BGL prediction using different approaches for data fusion of PA was evaluated and compared for two prediction horizons of 30 and 60 minutes. The evaluation was performed based on regression-wised criteria, including RMSE and MAE, and clinical-wised criteria, including MCC and SE, as presented in Section 2.4.1.

5.2.6 Statistical analyses

The performance of BGL prediction using various data fusion approaches was also statistically evaluated and compared over data contributors. Approaches for each level of PA fusion and the no-fusion approach were compared pair-wisely based on the recommended statistical tests in [86]. To do so, using a Friedman test [90], it was determined if there was a significant difference in the performance of BGL prediction between at least two approaches. Next, the Post-hoc Nemenyi test [92] was performed for pair-wise comparisons to determine which approaches performed significantly differently in a pair-wise fashion, with a significance level of 5%. Furthermore, the results of the post-hoc test were depicted by a CD diagram [86]. These analyses were then also performed between effective PA fusion approaches of each fusion level.

5.3 Results and discussion

This section presents the evaluation results of different PA fusion approaches along with rigorous statistical analyses for the two prediction horizons of 30 and 60 minutes. It is worth noting that since LSTM models rely on random initialisation, their performance was evaluated ten times, and the mean and standard deviation are reported for evaluation metrics. Also, when reporting statistical results, significant p-values with a significance level of 5% are highlighted in bold.

5.3.1 No-fusion

Tables 5.2 and 5.3 present the evaluation results of the BGL prediction using the no-fusion approach, in which BGL data was used as the only input, for prediction horizons of 30 and 60 minutes, respectively.

Table 5.2: Evaluation results of the BGL prediction using no-fusion approach for the prediction horizon of 30 minutes.

PID	RMSE	MAE	MCC	SE
559	19.85 ± 0.18	13.95 ± 0.22	0.79 ± 0.01	0.20 ± 0.00
563	18.80 ± 0.09	13.09 ± 0.08	0.77 ± 0.00	0.18 ± 0.00
570	23.50 ± 0.61	17.51 ± 0.63	0.82 ± 0.00	0.16 ± 0.00
575	24.08 ± 0.36	15.50 ± 0.52	0.73 ± 0.00	0.24 ± 0.01
588	18.88 ± 0.08	13.60 ± 0.05	0.74 ± 0.00	0.18 ± 0.00
591	22.68 ± 0.15	16.47 ± 0.10	0.64 ± 0.01	0.28 ± 0.00
Avg	21.30 ± 0.24	15.02 ± 0.27	0.75 ± 0.00	0.21 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Table 5.3: Evaluation results of the BGL prediction using no-fusion approach for the prediction horizon of 60 minutes.

PID	RMSE	MAE	MCC	SE
559	35.07 ± 0.11	25.88 ± 0.19	0.62 ± 0.01	0.35 ± 0.01
563	34.16 ± 1.85	25.46 ± 1.78	0.46 ± 0.06	0.35 ± 0.03
570	29.01 ± 0.37	21.11 ± 0.27	0.79 ± 0.01	0.20 ± 0.00
575	37.86 ± 0.28	27.83 ± 1.48	0.52 ± 0.02	0.43 ± 0.03
588	37.78 ± 3.99	28.46 ± 3.45	0.43 ± 0.07	0.38 ± 0.05
591	37.85 ± 0.80	29.95 ± 0.80	0.38 ± 0.01	0.47 ± 0.01
Avg	35.29 ± 1.23	26.45 ± 1.33	0.53 ± 0.03	0.36 ± 0.02

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

5.3.2 Signal-level PA fusion

Tables 5.4 and 5.5 show the results of evaluating the BGL prediction models that use BGL data fused with different signal-level information from PA to make predictions 30 and 60 minutes in advance, respectively.

To have a pair-wise comparison between the no-fusion approach and signal-level PA fusion approaches, first, the Friedman test was performed for both prediction horizons and all evaluation metrics. According to Table 5.6, there is sufficient

Table 5.4: Evaluation results of the BGL prediction using signal-level physical activity fusion approaches for the prediction horizon of 30 minutes.

PID	Input	RMSE	MAE	MCC	SE
559	BG+HR	19.98 ± 0.05	13.86 ± 0.13	0.81 ± 0.01	0.19 ± 0.00
	BG+HRSC	23.69 ± 0.61	16.07 ± 0.26	0.78 ± 0.01	0.21 ± 0.00
	BG+Band	23.41 ± 0.52	16.14 ± 0.23	0.77 ± 0.00	0.22 ± 0.00
563	BG+HR	18.94 ± 0.14	13.21 ± 0.12	0.77 ± 0.01	0.18 ± 0.00
	BG+HRSC	19.26 ± 0.19	13.47 ± 0.18	0.77 ± 0.01	0.19 ± 0.00
	BG+Band	19.41 ± 0.16	13.67 ± 0.10	0.77 ± 0.01	0.19 ± 0.00
570	BG+HR	16.55 ± 0.14	11.53 ± 0.11	0.87 ± 0.01	0.11 ± 0.00
	BG+HRSC	16.89 ± 0.52	11.67 ± 0.32	0.87 ± 0.01	0.11 ± 0.00
	BG+Band	17.93 ± 0.36	12.35 ± 0.35	0.85 ± 0.01	0.12 ± 0.00
575	BG+HR	24.00 ± 0.41	15.26 ± 0.25	0.74 ± 0.02	0.23 ± 0.01
	BG+HRSC	24.32 ± 0.13	15.55 ± 0.23	0.74 ± 0.02	0.23 ± 0.00
	BG+Band	24.45 ± 0.19	15.60 ± 0.16	0.74 ± 0.01	0.23 ± 0.00
588	BG+HR	18.92 ± 0.11	13.86 ± 0.29	0.75 ± 0.01	0.19 ± 0.01
	BG+HRSC	19.32 ± 0.34	13.98 ± 0.30	0.74 ± 0.02	0.19 ± 0.00
	BG+Band	19.48 ± 0.19	14.06 ± 0.17	0.73 ± 0.00	0.19 ± 0.00
591	BG+HR	22.64 ± 0.40	16.23 ± 0.39	0.64 ± 0.01	0.28 ± 0.01
	BG+HRSC	22.81 ± 0.39	16.38 ± 0.34	0.64 ± 0.01	0.28 ± 0.01
	BG+Band	22.95 ± 0.27	16.40 ± 0.23	0.65 ± 0.01	0.28 ± 0.00
Avg	BG+HR	20.17 ± 0.21	13.99 ± 0.22	0.76 ± 0.01	0.20 ± 0.00
	BG+HRSC	21.05 ± 0.36	14.52 ± 0.27	0.76 ± 0.01	0.20 ± 0.00
	BG+Band	21.27 ± 0.28	14.70 ± 0.21	0.75 ± 0.01	0.20 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; BG+HR, BG+HRSC, and BG+Band: Approaches for the signal-level fusion of physical activity data with blood glucose data.

evidence to be inferred that at least two approaches may perform differently for the BGL prediction. Therefore, in the next step, the post-hoc Nemenyi test was performed for pair-wise comparisons to determine which PA fusion approaches performed significantly differently. The results of the Nemenyi tests based on each evaluation metric are graphically represented as CD diagrams where horizontal lines link approaches with similar performances at a significance level of 5%. Then, to have an overview, CD diagrams according to the average ranking over all evaluation criteria were generated for each prediction horizon of 30 and 60 minutes. To be concise, individual CD diagrams related to each metric are presented in Figure 5.2 and CD diagrams based on the average over all metrics are presented in Figure 5.3.

Considering Figures 5.3a and 5.3b, which show the ranking for prediction horizon of 30 and 60 minutes, respectively, it can be concluded that BG+HR approach

Table 5.5: Evaluation results of the BGL prediction using signal-level physical activity fusion approaches for the prediction horizon of 60 minutes.

PID	Input	RMSE	MAE	MCC	SE
	BG+HR	35.17 ± 0.33	25.62 ± 0.30	0.63 ± 0.01	0.33 ± 0.01
559	BG+HRSC	39.13 ± 1.06	28.42 ± 0.88	0.61 ± 0.03	0.37 ± 0.02
	BG+Band	40.32 ± 1.13	29.02 ± 0.71	0.59 ± 0.01	0.38 ± 0.01
	BG+HR	31.40 ± 1.19	23.04 ± 1.17	0.54 ± 0.04	0.32 ± 0.02
563	BG+HRSC	31.67 ± 0.20	23.38 ± 0.25	0.55 ± 0.01	0.31 ± 0.00
	BG+Band	32.04 ± 0.39	23.31 ± 0.58	0.54 ± 0.02	0.32 ± 0.01
	BG+HR	28.58 ± 0.54	20.79 ± 0.45	0.79 ± 0.01	0.19 ± 0.00
570	BG+HRSC	28.54 ± 0.32	20.78 ± 0.23	0.78 ± 0.01	0.20 ± 0.00
	BG+Band	29.66 ± 0.47	21.51 ± 0.38	0.75 ± 0.00	0.21 ± 0.00
	BG+HR	37.71 ± 0.24	26.43 ± 0.17	0.53 ± 0.01	0.39 ± 0.00
575	BG+HRSC	38.55 ± 0.57	27.37 ± 0.45	0.52 ± 0.01	0.41 ± 0.01
	BG+Band	39.25 ± 0.39	28.02 ± 0.33	0.52 ± 0.01	0.41 ± 0.01
	BG+HR	32.31 ± 0.79	23.46 ± 0.40	0.56 ± 0.02	0.31 ± 0.00
588	BG+HRSC	32.42 ± 0.40	23.60 ± 0.33	0.56 ± 0.01	0.31 ± 0.00
	BG+Band	33.43 ± 0.51	24.44 ± 0.34	0.55 ± 0.00	0.31 ± 0.00
	BG+HR	37.10 ± 0.79	29.19 ± 0.98	0.39 ± 0.02	0.46 ± 0.01
591	BG+HRSC	37.00 ± 1.17	28.63 ± 1.31	0.41 ± 0.02	0.46 ± 0.02
	BG+Band	36.83 ± 0.91	28.84 ± 1.23	0.43 ± 0.03	0.46 ± 0.02
	BG+HR	33.71 ± 0.65	24.76 ± 0.58	0.57 ± 0.02	0.33 ± 0.01
Avg	BG+HRSC	34.55 ± 0.62	25.36 ± 0.58	0.57 ± 0.01	0.34 ± 0.01
	BG+Band	35.25 ± 0.63	25.86 ± 0.60	0.56 ± 0.01	0.35 ± 0.01

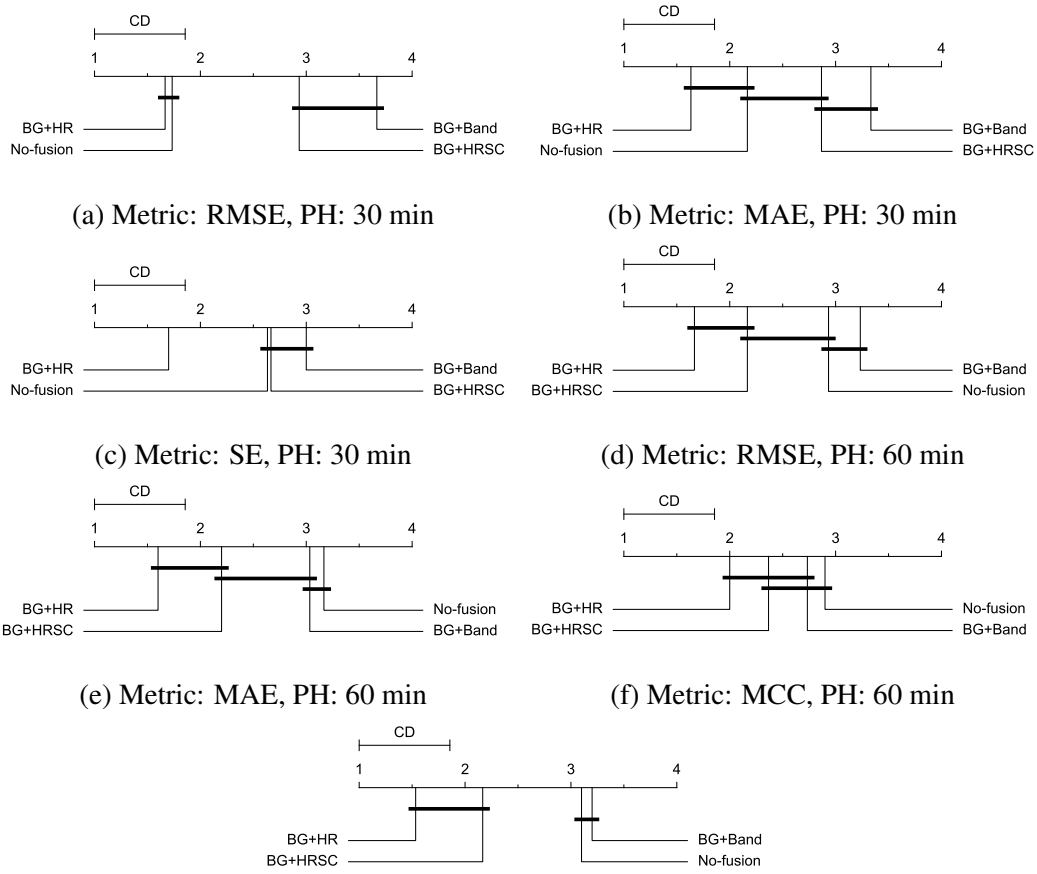
Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; BG+HR, BG+HRSC, and BG+Band: Approaches for the signal-level fusion of physical activity data with blood glucose data.

Table 5.6: p-values of the Friedman test for the comparison of BGL prediction performance using no-fusion approach and signal-level physical activity fusion approaches for prediction horizons of 30 and 60 minutes.

PH: 30 min				PH: 60 min			
RMSE	MAE	MCC	SE	RMSE	MAE	MCC	SE
0.000	**0.000**	0.058	**0.001**	**0.000**	**0.000**	**0.034**	**0.000**

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

was the best approach among signal-level PA fusion approaches and statistically significantly outperformed the no-fusion approach. Considering Tables 5.2, 5.3, 5.4, and 5.5 and Figure 5.2, it can be concluded that the BG+HR approach improved the



(a) Metric: RMSE, PH: 30 min

(b) Metric: MAE, PH: 30 min

(c) Metric: SE, PH: 30 min

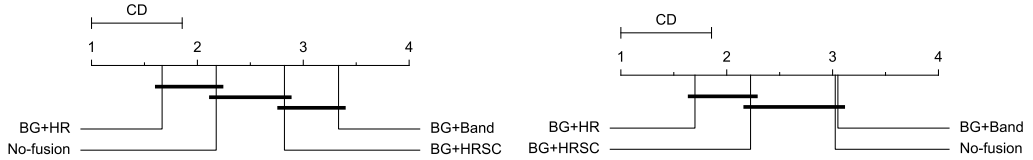
(d) Metric: RMSE, PH: 60 min

(e) Metric: MAE, PH: 60 min

(f) Metric: MCC, PH: 60 min

(g) Metric: SE, PH: 60 min

Figure 5.2: Critical difference diagram showing the comparison of the no-fusion approach and signal-level physical activity fusion approaches against each other according to RMSE (a), MAE (b), and SE (c) for the prediction horizon of 30 minutes as well as RMSE (d), MAE (e), MCC (f), and SE (g) for the prediction horizon of 60 minutes.



(a) PH: 30 min

(b) PH: 60 min

Figure 5.3: Critical difference diagram showing the comparison of the no-fusion approach and signal-level physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

average SE by 5.296% for the prediction horizon of 30 minutes. Also, for the prediction horizon of 60 minutes, this fusion approach improved RMSE, MAE, MCC,

and SE by 4.466%, 6.403%, 7.479% and 7.539%, respectively.

5.3.3 Feature-level PA fusion

The evaluation results of BGL prediction 30 and 60 minutes in advance using BGL data fused with different PA-driven features are presented in Tables 5.7 and 5.8.

Table 5.7: Evaluation results of the BGL prediction using feature-level physical activity fusion approaches for the prediction horizon of 30 minutes.

PID	Input	RMSE	MAE	MCC	SE
559	BG+SPA	20.28 ± 0.27	14.21 ± 0.15	0.81 ± 0.01	0.19 ± 0.00
	BG+OPA	21.39 ± 0.28	14.90 ± 0.12	0.79 ± 0.01	0.20 ± 0.00
	BG+StPA	21.14 ± 0.26	14.61 ± 0.07	0.80 ± 0.01	0.20 ± 0.00
563	BG+SPA	18.88 ± 0.07	13.17 ± 0.09	0.77 ± 0.00	0.18 ± 0.00
	BG+OPA	19.01 ± 0.08	13.32 ± 0.08	0.77 ± 0.01	0.18 ± 0.00
	BG+StPA	20.76 ± 0.28	14.42 ± 0.25	0.73 ± 0.01	0.20 ± 0.00
570	BG+SPA	16.40 ± 0.23	11.38 ± 0.10	0.87 ± 0.01	0.11 ± 0.00
	BG+OPA	17.02 ± 0.14	11.94 ± 0.11	0.87 ± 0.01	0.12 ± 0.00
	BG+StPA	17.33 ± 0.23	12.17 ± 0.19	0.87 ± 0.00	0.12 ± 0.00
575	BG+SPA	23.91 ± 0.20	15.30 ± 0.07	0.73 ± 0.00	0.22 ± 0.00
	BG+OPA	24.61 ± 0.41	15.46 ± 0.27	0.74 ± 0.01	0.23 ± 0.00
	BG+StPA	23.73 ± 0.48	14.97 ± 0.30	0.75 ± 0.01	0.23 ± 0.01
588	BG+SPA	18.53 ± 0.41	13.67 ± 0.27	0.77 ± 0.01	0.18 ± 0.01
	BG+OPA	19.02 ± 0.24	13.74 ± 0.19	0.75 ± 0.01	0.18 ± 0.00
	BG+StPA	18.70 ± 0.11	13.73 ± 0.23	0.76 ± 0.01	0.18 ± 0.00
591	BG+SPA	22.60 ± 0.15	16.45 ± 0.15	0.66 ± 0.00	0.28 ± 0.00
	BG+OPA	22.86 ± 0.42	16.30 ± 0.28	0.65 ± 0.00	0.28 ± 0.00
	BG+StPA	22.53 ± 0.30	16.34 ± 0.29	0.63 ± 0.02	0.28 ± 0.01
Avg	BG+SPA	20.10 ± 0.22	14.03 ± 0.14	0.77 ± 0.00	0.20 ± 0.00
	BG+OPA	20.65 ± 0.26	14.28 ± 0.17	0.76 ± 0.01	0.20 ± 0.00
	BG+StPA	20.70 ± 0.28	14.37 ± 0.22	0.75 ± 0.01	0.20 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; BG+SPA, BG+OPA, and BG+StPA: Approaches for the feature-level fusion of physical activity data with blood glucose data.

Table 5.9 shows the p-values of the Friedman test comparing the no-fusion approach and feature-level PA fusion approaches. Whenever the p-value for Friedman test was significant for any metric, a Nemenyi test was performed and visualised as CD diagrams. Similar to the previous section, CD diagrams related to each metric are shown in Figure 5.4, and CD diagrams based on average over all the significant metrics are visualised in Figure 5.5.

Table 5.8: Evaluation results of the BGL prediction using feature-level physical activity fusion approaches for the prediction horizon of 60 minutes.

PID	Input	RMSE	MAE	MCC	SE
559	BG+SPA	35.78 ± 0.48	26.10 ± 0.25	0.61 ± 0.01	0.34 ± 0.00
	BG+OPA	36.59 ± 0.15	26.55 ± 0.14	0.63 ± 0.02	0.34 ± 0.01
	BG+StPA	35.83 ± 0.52	26.14 ± 0.19	0.61 ± 0.02	0.35 ± 0.00
563	BG+SPA	31.40 ± 0.46	22.96 ± 0.59	0.55 ± 0.03	0.31 ± 0.01
	BG+OPA	31.46 ± 0.68	23.16 ± 0.58	0.55 ± 0.02	0.32 ± 0.01
	BG+StPA	32.98 ± 0.12	23.63 ± 0.32	0.51 ± 0.02	0.33 ± 0.01
570	BG+SPA	29.10 ± 1.03	21.01 ± 0.74	0.77 ± 0.01	0.20 ± 0.00
	BG+OPA	28.80 ± 0.34	21.12 ± 0.25	0.79 ± 0.01	0.20 ± 0.01
	BG+StPA	28.88 ± 0.22	21.08 ± 0.25	0.79 ± 0.01	0.19 ± 0.00
575	BG+SPA	37.96 ± 0.44	26.74 ± 0.23	0.50 ± 0.01	0.39 ± 0.00
	BG+OPA	37.98 ± 0.34	26.60 ± 0.25	0.52 ± 0.01	0.39 ± 0.00
	BG+StPA	37.82 ± 0.42	26.80 ± 0.19	0.52 ± 0.02	0.41 ± 0.01
588	BG+SPA	31.42 ± 0.22	22.68 ± 0.18	0.60 ± 0.01	0.29 ± 0.00
	BG+OPA	32.31 ± 0.24	23.29 ± 0.18	0.58 ± 0.01	0.30 ± 0.00
	BG+StPA	31.31 ± 0.34	22.76 ± 0.32	0.56 ± 0.01	0.30 ± 0.00
591	BG+SPA	36.65 ± 1.12	28.90 ± 1.15	0.45 ± 0.01	0.46 ± 0.02
	BG+OPA	36.70 ± 0.80	28.86 ± 0.89	0.39 ± 0.01	0.46 ± 0.01
	BG+StPA	35.28 ± 0.63	27.07 ± 0.68	0.42 ± 0.02	0.43 ± 0.01
Avg	BG+SPA	33.72 ± 0.63	24.73 ± 0.52	0.58 ± 0.01	0.33 ± 0.01
	BG+OPA	33.97 ± 0.42	24.93 ± 0.38	0.58 ± 0.01	0.33 ± 0.01
	BG+StPA	33.68 ± 0.37	24.58 ± 0.32	0.56 ± 0.01	0.34 ± 0.01

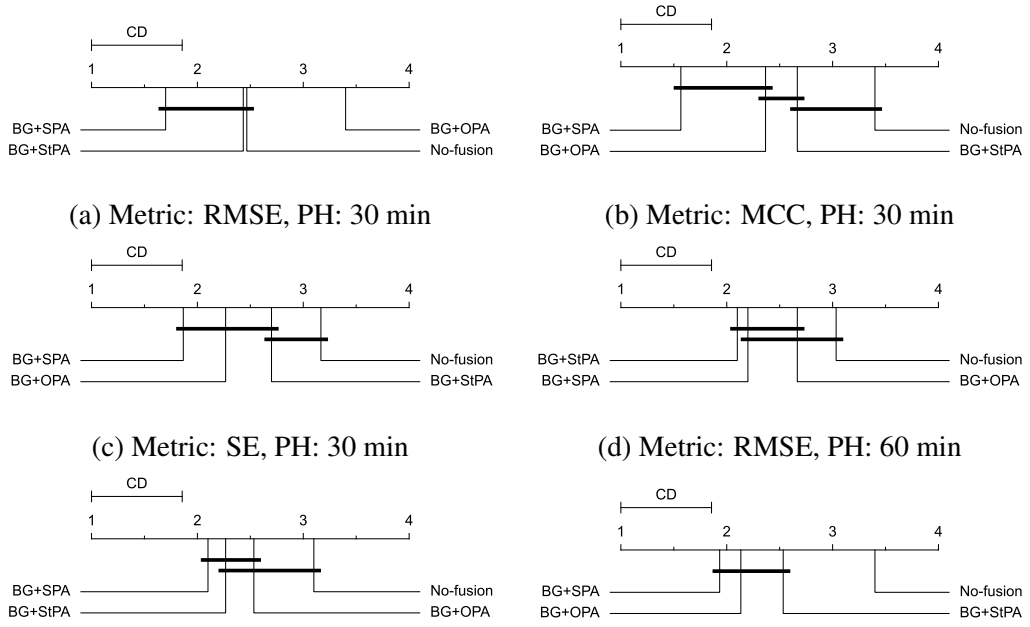
Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; BG+SPA, BG+OPA, and BG+StPA: Approaches for the feature-level fusion of physical activity data with blood glucose data.

Table 5.9: p-values of the Friedman test for the comparison of BGL prediction performance using no-fusion approach and feature-level physical activity fusion approaches for prediction horizons of 30 and 60 minutes.

PH: 30 min				PH: 60 min			
RMSE	MAE	MCC	SE	RMSE	MAE	MCC	SE
0.000	0.137	**0.000**	**0.003**	**0.000**	**0.001**	0.052	**0.000**

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

According to Figures 5.5a and 5.5b, it can be inferred that the BG+SPA approach was the best among feature-level PA fusion approaches, outperforming the no-fusion approach for both prediction horizons of 30 and 60 minutes. Considering



(a) Metric: RMSE, PH: 30 min

(b) Metric: MCC, PH: 30 min

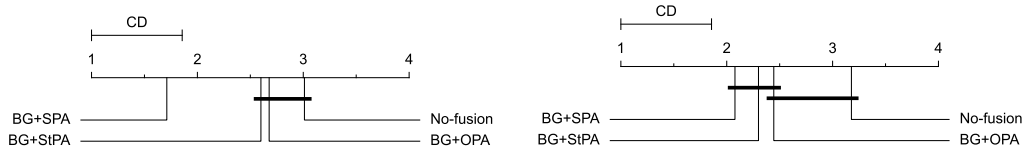
(c) Metric: SE, PH: 30 min

(d) Metric: RMSE, PH: 60 min

(e) Metric: MAE, PH: 60 min

(f) Metric: SE, PH: 60 min

Figure 5.4: Critical difference diagram showing the comparison of the no-fusion approach and feature-level physical activity fusion approaches against each other according to RMSE (a), MCC (b), and SE (c) for the prediction horizon of 30 minutes as well as RMSE (d), MAE (e), and SE (f) for the prediction horizon of 60 minutes.



(a) PH: 30 min

(b) PH: 60 min

Figure 5.5: Critical difference diagram showing the comparison of the no-fusion approach and feature-level physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

Tables 5.2, 5.3, 5.7, and 5.8 and Figure 5.4, it can be inferred that the BG+SPA approach improved the average evaluation metrics of MCC and SE over all patients by 2.921% and 6.063%, respectively, for the prediction horizon of 30 minutes, compared to the no-fusion approach. In addition, the BG+SPA improved the average values of MAE and SE by 6.495%, and 8.183%, respectively, compared to the no-fusion approach for the prediction horizon of 60 minutes.

5.3.4 Decision-level PA fusion

Tables 5.10 and 5.11 display the evaluation results of BGL prediction models that fuse decision-level information of PA and BGL using ensemble learning for prediction horizons of 30 and 60 minutes, respectively.

Table 5.10: Evaluation results of the BGL prediction using decision-level physical activity fusion approaches for the prediction horizon of 30 minutes.

PID	Input	RMSE	MAE	MCC	SE
559	BG&HR	19.49 ± 0.06	13.55 ± 0.05	0.79 ± 0.01	0.20 ± 0.00
	BG&HRSC	19.57 ± 0.07	13.63 ± 0.06	0.79 ± 0.01	0.20 ± 0.00
	BG&Band	19.60 ± 0.08	13.65 ± 0.06	0.79 ± 0.00	0.20 ± 0.00
563	BG&HR	18.71 ± 0.12	13.02 ± 0.10	0.76 ± 0.01	0.18 ± 0.00
	BG&HRSC	18.70 ± 0.13	13.00 ± 0.11	0.76 ± 0.01	0.18 ± 0.00
	BG&Band	18.68 ± 0.16	13.02 ± 0.11	0.77 ± 0.00	0.18 ± 0.00
570	BG&HR	17.38 ± 0.47	12.11 ± 0.32	0.86 ± 0.00	0.12 ± 0.00
	BG&HRSC	17.38 ± 0.48	12.11 ± 0.32	0.86 ± 0.00	0.12 ± 0.00
	BG&Band	17.39 ± 0.46	12.12 ± 0.31	0.86 ± 0.00	0.12 ± 0.00
575	BG&HR	24.03 ± 0.39	15.30 ± 0.25	0.74 ± 0.01	0.23 ± 0.01
	BG&HRSC	24.04 ± 0.40	15.31 ± 0.25	0.74 ± 0.01	0.24 ± 0.01
	BG&Band	24.03 ± 0.40	15.30 ± 0.25	0.74 ± 0.01	0.23 ± 0.01
588	BG&HR	19.13 ± 0.08	13.71 ± 0.08	0.74 ± 0.01	0.18 ± 0.00
	BG&HRSC	19.09 ± 0.10	13.74 ± 0.07	0.74 ± 0.01	0.18 ± 0.00
	BG&Band	19.31 ± 0.08	13.82 ± 0.09	0.74 ± 0.01	0.18 ± 0.00
591	BG&HR	22.33 ± 0.27	16.42 ± 0.28	0.62 ± 0.01	0.29 ± 0.00
	BG&HRSC	22.32 ± 0.24	16.42 ± 0.25	0.62 ± 0.01	0.29 ± 0.00
	BG&Band	22.53 ± 0.31	16.86 ± 0.34	0.62 ± 0.00	0.29 ± 0.00
Avg	BG&HR	20.18 ± 0.23	14.02 ± 0.18	0.75 ± 0.01	0.20 ± 0.00
	BG&HRSC	20.18 ± 0.24	14.03 ± 0.18	0.75 ± 0.01	0.20 ± 0.00
	BG&Band	20.26 ± 0.25	14.13 ± 0.19	0.75 ± 0.01	0.20 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; BG&HR, BG&HRSC, and BG&Band: Approaches for the decision-level fusion of physical activity data with blood glucose data.

The Friedman test was performed for comparison of the no-fusion and decision-level PA fusion approaches (Table 5.12). This was followed by the post-hoc Nemenyi test for significantly different metrics. CD diagrams visualising Nemenyi tests related to these metrics are shown in Figure 5.6, and CD diagrams based on average over all the metrics are visualised in Figure 5.7.

Based on Figures 5.7a and 5.7b, it can be concluded that for both prediction horizons of 30 and 60 minutes, the BG&HR approach outperformed the no-fusion

Table 5.11: Evaluation results of the BGL prediction using decision-level physical activity fusion approaches for the prediction horizon of 60 minutes.

PID	Input	RMSE	MAE	MCC	SE
559	BG&HR	34.37 ± 0.29	25.53 ± 0.19	0.62 ± 0.00	0.35 ± 0.00
	BG&HRSC	34.85 ± 0.24	26.02 ± 0.16	0.60 ± 0.01	0.35 ± 0.00
	BG&Band	34.93 ± 0.37	26.09 ± 0.30	0.60 ± 0.00	0.36 ± 0.00
563	BG&HR	32.75 ± 1.69	24.28 ± 1.74	0.48 ± 0.06	0.33 ± 0.03
	BG&HRSC	32.77 ± 1.70	24.29 ± 1.71	0.48 ± 0.06	0.33 ± 0.03
	BG&Band	32.62 ± 1.68	24.12 ± 1.61	0.49 ± 0.06	0.33 ± 0.02
570	BG&HR	28.70 ± 0.27	20.82 ± 0.21	0.79 ± 0.00	0.19 ± 0.00
	BG&HRSC	28.68 ± 0.27	20.82 ± 0.21	0.79 ± 0.00	0.19 ± 0.00
	BG&Band	28.74 ± 0.32	20.87 ± 0.25	0.79 ± 0.00	0.19 ± 0.00
575	BG&HR	37.35 ± 0.56	26.74 ± 0.52	0.51 ± 0.01	0.40 ± 0.01
	BG&HRSC	37.31 ± 0.57	26.69 ± 0.54	0.51 ± 0.01	0.40 ± 0.01
	BG&Band	37.30 ± 0.59	26.71 ± 0.56	0.51 ± 0.01	0.40 ± 0.01
588	BG&HR	38.26 ± 4.21	28.28 ± 3.47	0.44 ± 0.07	0.37 ± 0.04
	BG&HRSC	38.37 ± 4.21	28.43 ± 3.44	0.43 ± 0.07	0.37 ± 0.04
	BG&Band	38.50 ± 4.12	28.43 ± 3.39	0.45 ± 0.07	0.37 ± 0.04
591	BG&HR	37.35 ± 0.61	29.66 ± 0.57	0.36 ± 0.01	0.47 ± 0.01
	BG&HRSC	37.64 ± 0.64	29.94 ± 0.57	0.35 ± 0.01	0.47 ± 0.01
	BG&Band	38.16 ± 0.76	30.61 ± 0.60	0.34 ± 0.01	0.48 ± 0.01
Avg	BG&HR	34.80 ± 1.27	25.89 ± 1.12	0.53 ± 0.03	0.35 ± 0.02
	BG&HRSC	34.94 ± 1.27	26.03 ± 1.10	0.53 ± 0.03	0.36 ± 0.02
	BG&Band	35.04 ± 1.31	26.14 ± 1.12	0.53 ± 0.03	0.36 ± 0.01

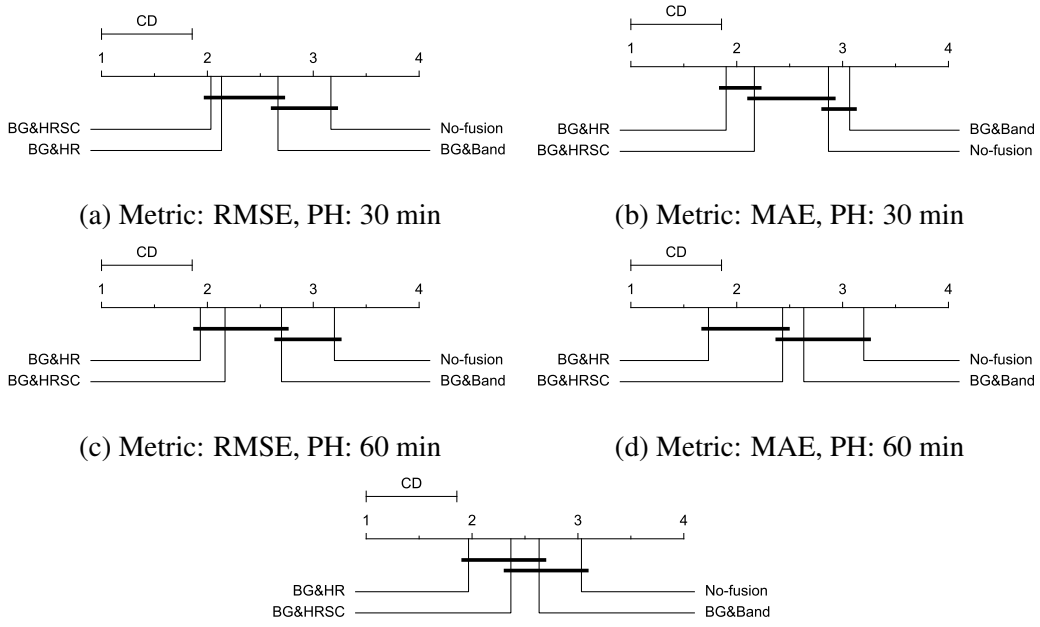
Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; BG&HR, BG&HRSC, and BG&Band: Approaches for the decision-level fusion of physical activity data with blood glucose data.

Table 5.12: p-values of the Friedman test for the comparison of BGL prediction performance using no-fusion approach and decision-level physical activity fusion approaches for prediction horizons of 30 and 60 minutes.

PH: 30 min				PH: 60 min			
RMSE	MAE	MCC	SE	RMSE	MAE	MCC	SE
0.002	**0.001**	0.199	0.106	**0.001**	**0.000**	0.215	**0.012**

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

approach. Considering Tables 5.2, 5.3, 5.10, and 5.11 and Figure 5.6, it can be inferred that the BG&HR approach improved the average evaluation metrics of RMSE and MAE over all patients by 5.265% and 6.684%, for the prediction horizon of 30



(a) Metric: RMSE, PH: 30 min

(b) Metric: MAE, PH: 30 min

(c) Metric: RMSE, PH: 60 min

(d) Metric: MAE, PH: 60 min

(e) Metric: SE, PH: 60 min

Figure 5.6: Critical difference diagram showing the comparison of the no-fusion approach and decision-level physical activity fusion approaches against each other according to RMSE (a) and MAE (b) for the prediction horizon of 30 minutes as well as RMSE (c), MAE (d), and SE (e) for the prediction horizon of 60 minutes.

(a) PH: 30 min

(b) PH: 60 min

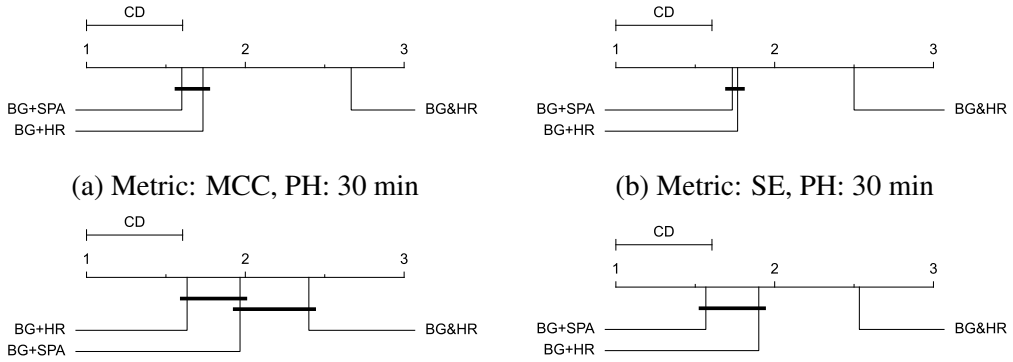
Figure 5.7: Critical difference diagram showing the comparison of the no-fusion approach and decision-level physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

minutes, respectively, compared to the no-fusion approach. Moreover, for the prediction horizon of 60 minutes, compared to the no-fusion approach, the BG&HR approach improved RMSE, MAE, and SE metrics by 1.394%, 2.128%, and 2.480%, respectively.

5.3.5 Comparison of the effective PA fusion approaches

As mentioned previously BG+HR, BG+SPA, and BG&HR approaches outperformed the no-fusion approach for at least one evaluation metric for both prediction hori-

zons. To compare these approaches with each other, a Friedman test was performed according to all evaluation metrics. According to the p-values of the Friedman test (Table 5.13), there was a significant difference between at least two PA fusion approaches regarding the MCC and SE evaluation metrics. Hence, the post-hoc Nemenyi test was performed on these metrics for pairwise comparisons. Similarly, CD diagrams visualising the outputs of Nemenyi tests based on each metric are shown in Figure 5.8. Also, CD diagrams based on the average over the two metrics are displayed in Figure 5.9.



(a) Metric: MCC, PH: 30 min

(b) Metric: SE, PH: 30 min

(c) Metric: MCC, PH: 60 min

(d) Metric: SE, PH: 60 min

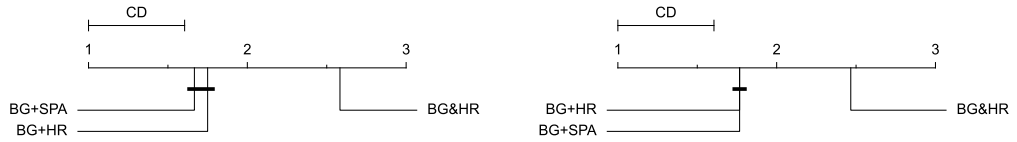
Figure 5.8: Critical difference diagram showing the comparison of the effective physical activity fusion approaches against each other according to MCC (a) and SE (b) for the prediction horizon of 30 minutes as well as MCC (c) and SE (d) for the prediction horizon of 60 minutes.

Table 5.13: p-values of the Friedman test for comparing the effective physical activity fusion approaches from different levels for prediction horizons of 30 and 60 minutes.

PH: 30 min				PH: 60 min			
RMSE	MAE	MCC	SE	RMSE	MAE	MCC	SE
0.393	0.967	**0.000**	**0.004**	0.531	0.150	**0.012**	**0.001**

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Considering Figure 5.9, it can be concluded that BG+HR and BG+SPA approaches similarly performed better than BG&HR approach.



(a) PH: 30 min (b) PH: 60 min

Figure 5.9: Critical difference diagram showing the comparison of the effective physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

5.4 Summary

The goal of this work was to contribute to finding optimal approaches for PA deployment in BGL prediction models, including the kind of PA information and the level of integration. This work developed different PA-informed models for BGL prediction by extracting various information from PA data and fusing this information with BGL data in signal, feature, and decision levels. To do so, for the signal-level fusion, three different combinations of automatically-recorded PA data from wristbands including (HR), (HR and SC), and (HR, SC, GSR, and ST) were fused with BGL data. Also, three feature engineering approaches including subjective PA categorisation, objective PA clustering, and statistics of PA were used to fuse feature-level PA information with BGL data. Moreover, for decision-level PA fusion, the primary decisions made by the base-learner using BGL data and PA data, separately, were stacked and used as inputs of the meta-learner. Based on the kind of PA data, three different decision-level approaches were developed. In total, nine PA fusion approaches were developed. These approaches were compared with the no-fusion approach and also with each other.

All in all, the results of the comparison of the PA-informed models with the no-fusion approach showed that fusing PA information with BGL can statistically significantly improve the performance of BGL prediction. Among all the developed PA fusion approaches, fusing BGL with the automatically recorded HR data and with categories of self-reported PA-related events outperformed the no-fusion and other PA fusion approaches.

Chapter 6

Benchmark of data-driven approaches for blood glucose level prediction

6.1 Preface

Comparing the performance of different prediction approaches can provide beneficial insight into BGL prediction. Also, using different datasets or input features in the literature has made the performance comparison of different models difficult. Hence, making fair comparisons is valuable research in BGL prediction [39]. Limited studies [39, 38] have been done in this investigation. Also, regarding input, there is some evidence that BGL prediction from BGL data alone facilitates practical application in the real world, therefore, suggesting that there is no need for the extra effort and cost to acquire and process data from several sensors and modalities [56, 71, 72, 73, 74, 75, 76]. Conversely, there is evidence that other variables can also contribute to the performance of BGL prediction [57, 77, 78]. Limited work [40, 41] has been performed for comparing different inputs in BGL prediction models. Previous studies have not provided an in-depth and comprehensive comparison of different prediction approaches or inputs. In addition, in the previous studies, the average prediction performance across the data providers was considered for the purpose of comparison. However, due to considerable variation between patients regarding BGL, this type of comparison would not be meaningful and for a more valid comparison, statistical analyses need to be considered. Hence, due to the lack of statistical analyses in these previous studies, their conclusions may not be robust.

The work presented in this chapter benchmarks BGL prediction from two perspectives; models' approaches and models' inputs. First, it compares the perfor-

mance of BGL prediction using different data-driven time series forecasting approaches, including classical time series forecasting, traditional machine learning, and deep neural networks. Secondly, a comparison between using BGL data only as a univariate input is compared to a multivariate input using BGL data in addition to data on carbohydrate intake, injected bolus insulin doses, and activity levels. This investigation demonstrates how adding exogenous variables impacts different time series forecasting approaches in the BGL prediction task. Regression-wised and clinical-wised metrics along with statistical analyses were performed for evaluation and comparison purposes.

6.2 Material and Methods

6.2.1 Dataset

This work used Ohio_2018 [66] and Ohio_2020 [67] datasets described in Section 2.8. In this work, BGL, carbohydrate, bolus, and PA data (HR for the Ohio_2018 dataset and MA for the Ohio_2020 dataset) were used.

6.2.2 Preprocessing

There were some mandatory preprocessing steps to overcome many imperfections and missing data when analysing real world data. Additionally, some data preprocessing was required dependent on the forecasting approach used.

6.2.2.1 Imputation and alignment

The initial preprocessing step was to address the issue of missing BGL and PA data. These missing values were interpolated in training and extrapolated in testing sets linearly. No reported timestamps for carbohydrate and bolus data were assigned to zero. The following preprocessing step was to align the BGL data with other data. Data of MA, with a resolution of one minute, was downsampled to a resolution of five minutes by taking the nearest MA data point with a BGL data point and removing the remainder. The HR data, which had the same resolution as BGL data, only required to be aligned. Additionally, the unavailable data timestamps at the beginning/ending of each set, which occurred due to different times in the wearing sensors, were discarded.

6.2.2.2 Stationarity

When applying the CTF approach, two common statistical tests were applied to check the primary assumption of stationarity [5]; the Augmented Dickey-Fuller (ADF) test [165] and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [166]. Time series in which both tests confirmed the stationarity were defined as stationary. Since, the ADF test indicated stationarity for all variables and all patients, integrated differencing was applied to the time series in which the KPSS indicated non-stationarity.

6.2.2.3 Reframing

When applying TML or DNN approaches, the multi-ahead time series forecasting problem should be reframed as a supervised learning task. To accomplish this, time series data were transformed into samples using historical observations as inputs, and future observations as outputs with a one-step rolling window. Also, throughout the entire training sets of all subjects, input sequences were scaled to minimum and maximum values.

6.2.3 Time series forecasting approaches

Data-driven models can be classified into classical time series forecasting (CTF), traditional machine learning (TML), and deep neural networks (DNNs) approaches. Comparing the efficacy of various data-driven prediction models using different approaches would be beneficial in the advancement of BGL prediction performance. To comprehensively investigate and compare the performance of BGL prediction, different time series forecasting categories were examined. There are a pool of models for BGL prediction in each category. For the sake of feasibility and in order to minimise the complexity of comparison, for each category, a common successful model found in the literature was developed and fine-tuned as a representative. For input comparison purposes, each model was first trained as a univariate prediction model; then, its counterpart was developed as a multivariate prediction model. The prediction models are briefly described in the following.

6.2.3.1 Classical time series forecasting

CTF is a common approach for the BGL prediction task [39, 49]. One of the most commonly used models in this category for univariate time series forecasting is ARIMA [52]. ARIMA is a combination of linear processes of AR and MA models, as well as integrated differencing. It models the future as a linear combination

of lags and lagged residual errors in a differenced time series in the case of non-stationarity. To develop an ARIMA model, the parameters of the model, including p (AR order), d (differencing order), and q (MA order), should be determined. The p and q parameters were optimised for each patient automatically by examining each parameter from zero to 36. The d parameter was also determined by considering the stationarity tests. An ARIMAX was used for the multivariate prediction, incorporating exogenous variables into the univariate ARIMA model. Table 6.1 shows the optimised parameters for each patient training the ARIMA and ARIMAX models.

Table 6.1: The optimised parameters for the ARIMA and ARIMAX models.

Ohio_2018				Ohio_2020			
PID	p	d	q	PID	p	d	q
559	6	0	2	540	4	1	2
563	3	0	2	544	5	1	3
570	3	0	2	552	3	1	1
575	4	1	3	567	1	1	2
588	1	1	1	584	2	0	3
591	2	0	4	596	3	1	1

Note. PID: Patient identity.

6.2.3.2 Traditional machine learning

A TML approach has also received significant attention for predicting BGL. SVMs have been shown to be the most accurate in the BGL prediction task among different classes of machine learning algorithms [19, 167]. Also, among different types of SVMs, SVR is the most commonly employed technique for predicting BGL [19]. In this study, in line with the successfully developed SVM model for BGL prediction in the literature [168], an SVR model with a radial basis kernel was developed. Moreover, to have a multivariate prediction using SVM, vectorised multivariate data were utilised as the input for developing multivariate counterparts. The hyperparameters of the SVR model, including gamma, C, and epsilon, were chosen using a grid search during a tuning process for each patient and each input. Search spaces of {0.1,1, 10, 100}, {0.001,0.01, 0.1, 1}, and {0.01, 0.1, 1, 10} were explored to optimise gamma, C, and epsilon parameters, respectively. The chosen parameters are summarised in Table 6.2.

Table 6.2: The optimised parameters for the SVR model.

	PID	Univariate						Multivariate					
		PH:30 min			PH:60 min			PH:30 min			PH:60 min		
		γ	c	ϵ	γ	c	ϵ	γ	c	ϵ	γ	c	ϵ
Ohio_2018	559	100	10	1	100	10	1	100	10	0.1	100	10	0.1
	563	100	10	0.1	100	10	0.1	100	0.01	0.1	100	10	0.1
	570	100	1	1	10	1	1	100	1	0.1	100	10	0.1
	575	100	0.01	1	100	10	1	100	0.01	0.1	100	10	0.1
	588	100	10	0.1	100	10	1	100	1	0.1	100	10	0.1
	591	100	10	1	10	0.01	1	100	10	0.1	100	10	0.1
Ohio_2020	540	100	10	1	100	10	1	100	10	0.1	100	10	0.1
	544	100	10	1	100	10	1	100	1	0.01	100	10	0.1
	552	100	10	1	100	10	1	100	10	0.1	100	10	0.1
	567	100	10	1	100	10	1	100	10	0.1	100	10	1
	584	100	10	1	100	10	1	100	10	0.1	100	10	0.1
	596	100	10	1	100	1	0.1	100	10	0.1	100	10	1

Note. PID: Patient identity; PH: Prediction horizon.

6.2.3.3 Deep neural networks

LSTM networks, as a type of recurrent neural networks, are effective at predicting BGL based on sequential data [83, 95, 117, 118]. In this study, the sequence-to-sequence forecasting task was carried out using the LSTM model developed and discussed in Sections 3.2.3.2 and 4.2.4.1, which has been optimised in the Ohio datasets. The vanilla LSTM network consisted of a 200-unit LSTM layer, a 100-unit dense layer, and an output layer. The initialiser of He uniform, the activation function of ReLU, the optimiser of Adam, and the loss function of MSE were chosen. Also, an epoch size of 200 and a batch size of 32 were selected. The learning rate with an initial value of 0.01 was reduced by a factor of 0.1 following the usage of a ReduceLROnPlateau callback with a patience of 20 after stopping validation loss improvement.

6.2.4 Evaluation criteria

To comprehensively investigate the performance of BGL prediction using different prediction approaches and inputs regression-wised and clinical-wised evaluations were performed. To do so, RMSE and MAE, as regression-wised criteria, were used. Also, to evaluate the overall clinical performance MCC and SE were utilised. These metrics were calculated as described in Section 2.4.1.

6.2.5 Statistical analyses

The BGL prediction performance measured by evaluation metrics with various prediction approaches or inputs was also statistically analysed over data contributors for each dataset. In accordance with the conditions of each comparison, appropriate statistical analyses were conducted.

To compare different prediction models, firstly, the Friedman test [90] was conducted in order to determine whether there was a significant difference between at least two approaches (with a significance level of five percent). If this was the case, a post-hoc Nemenyi test [92] was then performed to compare the performance of different approaches in a pair-wise fashion. Also, since multiple comparisons were made, the Holm procedure [169] was applied to correct the significance level. A CD diagram [86] was drawn to illustrate the results of each post-hoc test. These analyses were performed for each univariate and multivariate input separately.

To compare univariate and multivariate inputs for each prediction approach, a non-parametric Wilcoxon signed-ranks test [89] was applied. This test, with a significance level of five percent was conducted to check the consistency of each evaluation metric calculated for univariate and multivariate inputs over the data contributors of each dataset. The comparison of input was performed for each prediction approach separately.

6.3 Results and discussion

In this section, firstly the results of evaluation criteria for both Ohio_2018 and Ohio_2020 datasets for both prediction horizons of 30 and 60 minutes are presented. Then, depending on which factor is being compared, results of relative statistical analyses are presented and discussed in two parts; comparing models' approaches and models' inputs.

6.3.1 Evaluation results

Tables 6.3 and 6.4 provide the results of evaluation criteria for the BGL prediction models related to different approaches for both univariate and multivariate inputs, 30 and 60 minutes in advance in Ohio_2018 dataset, respectively. Also, Tables 6.5 and 6.6 provide the evaluation results in Ohio_2020 dataset, for prediction horizons of 30 and 60 minutes, respectively. It is worth noting that for the DNN approach, due to the random initialization, the models were run 10 times. The average and standard deviation of evaluation results over 10 runs are reported. Using evaluation

results, to compare different models and inputs, statistical analyses were performed. The results are discussed in the following sections.

Table 6.3: Evaluation results of different prediction approaches and inputs in Ohio_2018 dataset for the prediction horizon of 30 minutes.

PID	Model	Input	RMSE	MAE	MCC	SE
559	CTF	univariate	20.07	13.82	0.78	0.19
		multivariate	20.12	13.86	0.79	0.20
	TML	univariate	20.56	14.00	0.81	0.19
		multivariate	19.35	13.34	0.83	0.18
	DNN	univariate	20.19 ± 0.18	14.16 ± 0.13	0.78 ± 0.01	0.21 ± 0.01
		multivariate	20.70 ± 0.41	14.68 ± 0.31	0.80 ± 0.01	0.20 ± 0.01
563	CTF	univariate	20.14	13.82	0.75	0.20
		multivariate	20.33	13.98	0.75	0.20
	TML	univariate	18.67	13.28	0.75	0.19
		multivariate	18.52	12.89	0.77	0.18
	DNN	univariate	18.93 ± 0.10	13.12 ± 0.13	0.77 ± 0.01	0.18 ± 0.00
		multivariate	20.45 ± 0.32	14.11 ± 0.24	0.76 ± 0.01	0.19 ± 0.00
570	CTF	univariate	17.01	12.17	0.86	0.12
		multivariate	17.15	12.32	0.85	0.12
	TML	univariate	17.24	11.71	0.87	0.11
		multivariate	16.09	11.20	0.87	0.10
	DNN	univariate	17.11 ± 0.52	11.97 ± 0.45	0.87 ± 0.01	0.11 ± 0.00
		multivariate	18.10 ± 0.40	12.58 ± 0.24	0.86 ± 0.01	0.12 ± 0.00
575	CTF	univariate	25.17	15.58	0.76	0.23
		multivariate	25.17	15.58	0.76	0.23
	TML	univariate	24.08	14.93	0.74	0.22
		multivariate	24.08	14.93	0.76	0.22
	DNN	univariate	24.42 ± 0.21	15.72 ± 0.24	0.73 ± 0.01	0.24 ± 0.01
		multivariate	25.79 ± 0.49	15.78 ± 0.39	0.72 ± 0.01	0.23 ± 0.01
588	CTF	univariate	19.62	14.19	0.74	0.19
		multivariate	19.62	14.20	0.74	0.19
	TML	univariate	21.28	15.34	0.69	0.20
		multivariate	18.03	13.09	0.75	0.17

(continued on next page)

Table 6.3 (continued)

PID	Model	Input	RMSE	MAE	MCC	SE
	DNN	univariate	18.84 ± 0.10	13.54 ± 0.07	0.75 ± 0.01	0.18 ± 0.00
		multivariate	18.84 ± 0.35	13.80 ± 0.34	0.76 ± 0.01	0.18 ± 0.00
591	CTF	univariate	22.65	16.03	0.66	0.27
		multivariate	22.69	16.06	0.65	0.27
	TML	univariate	21.78	15.61	0.65	0.27
		multivariate	21.49	15.50	0.65	0.26
	DNN	univariate	22.87 ± 0.45	16.59 ± 0.48	0.63 ± 0.01	0.29 ± 0.01
		multivariate	22.79 ± 0.31	16.47 ± 0.27	0.64 ± 0.01	0.28 ± 0.01
Avg	CTF	univariate	20.78	14.27	0.76	0.20
		multivariate	20.85	14.33	0.76	0.20
	TML	univariate	20.60	14.14	0.75	0.20
		multivariate	19.59	13.49	0.77	0.19
	DNN	univariate	20.39 ± 0.26	14.18 ± 0.25	0.75 ± 0.01	0.20 ± 0.00
		multivariate	21.11 ± 0.38	14.57 ± 0.30	0.76 ± 0.01	0.20 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CTF: Classical time series forecasting; TML: Traditional machine learning; DNN: Deep neural network.

Table 6.4: Evaluation results of different prediction approaches and inputs in Ohio_2018 dataset for the prediction horizon of 60 minutes.

PID	Model	Input	RMSE	MAE	MCC	SE
559	CTF	univariate	36.03	25.76	0.58	0.36
		multivariate	36.24	26.00	0.58	0.36
	TML	univariate	35.69	25.44	0.63	0.33
		multivariate	31.69	22.51	0.69	0.29
	DNN	univariate	35.83 ± 0.45	26.31 ± 0.27	0.62 ± 0.01	0.35 ± 0.01
		multivariate	35.52 ± 0.80	26.02 ± 0.74	0.61 ± 0.02	0.35 ± 0.01
563	CTF	univariate	33.01	24.39	0.54	0.34
		multivariate	32.84	24.36	0.53	0.34
	TML	univariate	30.32	22.13	0.54	0.31
		multivariate	30.32	21.72	0.59	0.29

(continued on next page)

Table 6.4 (continued)

PID	Model	Input	RMSE	MAE	MCC	SE
	DNN	univariate	32.25 ± 1.22	23.45 ± 1.33	0.52 ± 0.05	0.32 ± 0.02
		multivariate	33.63 ± 0.67	23.97 ± 0.54	0.54 ± 0.02	0.32 ± 0.01
570	CTF	univariate	30.20	22.84	0.75	0.22
		multivariate	30.37	23.01	0.74	0.22
	TML	univariate	29.50	21.17	0.79	0.19
		multivariate	27.67	19.98	0.79	0.18
	DNN	univariate	29.02 ± 0.62	20.75 ± 0.62	0.80 ± 0.00	0.19 ± 0.00
		multivariate	30.95 ± 0.46	22.23 ± 0.59	0.80 ± 0.01	0.20 ± 0.00
575	CTF	univariate	39.96	27.51	0.56	0.41
		multivariate	39.97	27.51	0.56	0.41
	TML	univariate	37.09	25.98	0.51	0.39
		multivariate	36.01	25.24	0.56	0.37
	DNN	univariate	38.09 ± 0.30	27.10 ± 0.18	0.50 ± 0.01	0.41 ± 0.00
		multivariate	40.02 ± 0.69	27.60 ± 0.29	0.51 ± 0.01	0.41 ± 0.00
588	CTF	univariate	33.98	25.15	0.57	0.33
		multivariate	33.98	25.16	0.57	0.33
	TML	univariate	31.43	22.73	0.56	0.29
		multivariate	30.21	22.28	0.59	0.28
	DNN	univariate	31.62 ± 0.16	23.24 ± 0.15	0.54 ± 0.01	0.31 ± 0.00
		multivariate	31.91 ± 0.42	23.31 ± 0.34	0.58 ± 0.02	0.30 ± 0.00
591	CTF	univariate	36.94	27.53	0.36	0.46
		multivariate	36.98	27.57	0.35	0.46
	TML	univariate	33.58	25.40	0.45	0.41
		multivariate	33.33	25.42	0.41	0.41
	DNN	univariate	36.71 ± 0.80	28.77 ± 0.78	0.38 ± 0.02	0.46 ± 0.01
		multivariate	35.69 ± 0.79	27.53 ± 0.67	0.44 ± 0.02	0.44 ± 0.01
Avg	CTF	univariate	35.02	25.53	0.56	0.35
		multivariate	35.06	25.60	0.56	0.36
	TML	univariate	32.93	23.81	0.58	0.32
		multivariate	31.54	22.86	0.61	0.30
	DNN	univariate	33.92 ± 0.59	24.94 ± 0.55	0.56 ± 0.02	0.34 ± 0.01

(continued on next page)

Table 6.4 (continued)

PID	Model	Input	RMSE	MAE	MCC	SE
		multivariate	34.62 ± 0.64	25.11 ± 0.53	0.58 ± 0.01	0.34 ± 0.01

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CTF: Classical time series forecasting; TML: Traditional machine learning; DNN: Deep neural network.

Table 6.5: Evaluation results of different prediction approaches and inputs in Ohio_2020 dataset for the prediction horizon of 30 minutes.

PID	Model	Input	RMSE	MAE	MCC	SE
540	CTF	univariate	21.46	16.13	0.73	0.25
		multivariate	22.01	16.24	0.74	0.25
	TML	univariate	29.07	18.34	0.71	0.26
		multivariate	23.11	16.83	0.71	0.26
	DNN	univariate	22.58 ± 0.77	16.82 ± 0.45	0.71 ± 0.01	0.25 ± 0.01
		multivariate	21.99 ± 0.89	16.33 ± 0.33	0.70 ± 0.01	0.26 ± 0.00
544	CTF	univariate	18.93	13.42	0.77	0.19
		multivariate	18.94	13.42	0.77	0.19
	TML	univariate	18.11	12.98	0.79	0.19
		multivariate	18.74	13.32	0.78	0.19
	DNN	univariate	18.14 ± 0.12	12.90 ± 0.13	0.79 ± 0.00	0.19 ± 0.00
		multivariate	19.04 ± 0.19	13.07 ± 0.13	0.78 ± 0.01	0.19 ± 0.00
552	CTF	univariate	17.42	12.30	0.74	0.21
		multivariate	17.42	12.30	0.74	0.21
	TML	univariate	17.01	12.47	0.74	0.21
		multivariate	16.88	12.88	0.70	0.23
	DNN	univariate	16.89 ± 0.05	12.49 ± 0.10	0.74 ± 0.01	0.21 ± 0.00
		multivariate	18.48 ± 0.77	13.55 ± 0.54	0.70 ± 0.02	0.23 ± 0.01
567	CTF	univariate	22.39	15.53	0.71	0.24
		multivariate	22.39	15.53	0.71	0.24
	TML	univariate	21.06	14.84	0.67	0.25
		multivariate	21.82	15.38	0.62	0.26
	DNN	univariate	21.22 ± 0.21	15.11 ± 0.23	0.65 ± 0.01	0.26 ± 0.00

(continued on next page)

Table 6.5 (continued)

PID	Model	Input	RMSE	MAE	MCC	SE
		multivariate	20.87 ± 0.30	14.67 ± 0.23	0.65 ± 0.02	0.25 ± 0.00
584	CTF	univariate	22.53	16.06	0.74	0.22
		multivariate	23.36	16.81	0.73	0.23
	TML	univariate	21.88	15.84	0.77	0.22
		multivariate	21.23	15.40	0.78	0.21
	DNN	univariate	23.16 ± 0.50	17.02 ± 0.43	0.76 ± 0.01	0.23 ± 0.00
		multivariate	22.66 ± 0.59	16.56 ± 0.46	0.77 ± 0.01	0.23 ± 0.01
596	CTF	univariate	18.88	13.50	0.71	0.22
		multivariate	18.88	13.50	0.71	0.22
	TML	univariate	17.89	12.76	0.74	0.21
		multivariate	16.86	12.21	0.78	0.19
	DNN	univariate	18.17 ± 0.11	12.94 ± 0.10	0.75 ± 0.01	0.21 ± 0.00
		multivariate	18.52 ± 0.38	13.11 ± 0.26	0.75 ± 0.01	0.21 ± 0.00
Avg	CTF	univariate	20.27	14.49	0.73	0.22
		multivariate	20.50	14.63	0.73	0.23
	TML	univariate	20.83	14.54	0.74	0.22
		multivariate	19.77	14.34	0.73	0.22
	DNN	univariate	20.03 ± 0.30	14.55 ± 0.24	0.73 ± 0.01	0.23 ± 0.00
		multivariate	20.26 ± 0.52	14.55 ± 0.33	0.72 ± 0.01	0.23 ± 0.00

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CTF: Classical time series forecasting; TML: Traditional machine learning; DNN: Deep neural network.

Table 6.6: Evaluation results of different prediction approaches and inputs in Ohio_2020 dataset for the prediction horizon of 60 minutes.

PID	Model	Input	RMSE	MAE	MCC	SE
540	CTF	univariate	40.42	31.13	0.52	0.46
		multivariate	42.54	32.46	0.51	0.48
	TML	univariate	44.81	32.49	0.50	0.45
		multivariate	41.42	30.90	0.54	0.44
	DNN	univariate	40.83 ± 1.33	30.98 ± 0.45	0.53 ± 0.01	0.44 ± 0.00

(continued on next page)

Table 6.6 (continued)

PID	Model	Input	RMSE	MAE	MCC	SE
		multivariate	41.75 ± 0.88	31.00 ± 0.48	0.53 ± 0.03	0.44 ± 0.01
544	CTF	univariate	34.84	25.36	0.57	0.36
		multivariate	34.85	25.35	0.57	0.36
	TML	univariate	32.01	23.42	0.61	0.33
		multivariate	28.25	20.49	0.66	0.30
	DNN	univariate	32.00 ± 0.21	24.69 ± 0.32	0.60 ± 0.01	0.36 ± 0.01
		multivariate	32.33 ± 1.07	22.74 ± 0.69	0.64 ± 0.02	0.33 ± 0.01
552	CTF	univariate	32.13	22.61	0.57	0.37
		multivariate	32.13	22.61	0.57	0.37
	TML	univariate	29.76	21.49	0.58	0.34
		multivariate	28.87	21.87	0.58	0.35
	DNN	univariate	30.32 ± 0.13	22.71 ± 0.17	0.58 ± 0.01	0.36 ± 0.00
		multivariate	30.98 ± 0.65	23.47 ± 0.54	0.56 ± 0.02	0.37 ± 0.01
567	CTF	univariate	42.34	30.13	0.48	0.46
		multivariate	42.34	30.13	0.48	0.46
	TML	univariate	37.16	27.31	0.40	0.44
		multivariate	37.46	27.40	0.44	0.44
	DNN	univariate	39.23 ± 1.86	30.28 ± 2.12	0.36 ± 0.02	0.51 ± 0.04
		multivariate	36.63 ± 0.13	27.42 ± 0.22	0.38 ± 0.01	0.47 ± 0.00
584	CTF	univariate	38.93	28.07	0.56	0.37
		multivariate	39.92	28.84	0.56	0.38
	TML	univariate	36.77	27.11	0.63	0.35
		multivariate	33.89	25.28	0.63	0.34
	DNN	univariate	39.83 ± 1.96	30.16 ± 1.78	0.59 ± 0.03	0.40 ± 0.02
		multivariate	38.38 ± 1.71	29.40 ± 1.76	0.57 ± 0.03	0.40 ± 0.02
596	CTF	univariate	33.20	24.29	0.51	0.38
		multivariate	33.20	24.28	0.51	0.38
	TML	univariate	30.27	22.18	0.57	0.33
		multivariate	27.82	20.15	0.61	0.30
	DNN	univariate	30.20 ± 0.21	22.22 ± 0.25	0.58 ± 0.02	0.33 ± 0.01
		multivariate	30.38 ± 1.07	22.45 ± 0.98	0.57 ± 0.03	0.33 ± 0.01

(continued on next page)

Table 6.6 (continued)

PID	Model	Input	RMSE	MAE	MCC	SE
Avg	CTF	univariate	36.98	26.93	0.53	0.40
		multivariate	37.50	27.28	0.53	0.40
	TML	univariate	35.13	25.67	0.55	0.38
		multivariate	32.95	24.35	0.58	0.36
	DNN	univariate	35.40 ± 0.95	26.84 ± 0.85	0.54 ± 0.01	0.40 ± 0.01
		multivariate	35.07 ± 0.92	26.08 ± 0.78	0.54 ± 0.02	0.39 ± 0.01

Note. PID: Patient identity; PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error; CTF: Classical time series forecasting; TML: Traditional machine learning; DNN: Deep neural network.

6.3.2 Comparing models' approaches

Different data-driven time series forecasting approaches are compared using univariate and multivariate inputs, separately. Firstly, the results of statistical analyses are presented and discussed. Secondly, computational costs for different models are compared. Then, according to all presented results, a conclusion is presented.

6.3.2.1 Statistical result

Univariate input Table 6.7 presents p-values of the Friedman test calculated based on evaluation metrics of BGL prediction performance using different forecasting approaches with a univariate input. The analysis was performed for both prediction horizons of 30 and 60 minutes, and for both Ohio_2018 and Ohio_2020 datasets, separately. With a significance level of five percent, p-values marked with bold font are related to the cases with probably at least one significant difference between the performance of models.

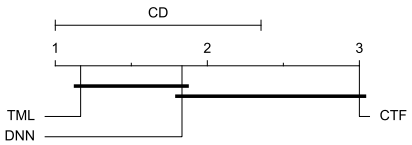
Reviewing Tables 6.3, 6.4, 6.5, 6.6, and 6.7, it can be concluded that although there are differences between average evaluation metrics related to the performance of different prediction models over data providers of each cohort, these differences are mainly statistically insignificant. Table 6.7 shows that just three metrics of RMSE, MAE, and SE calculated for a prediction horizon of 60 minutes in the Ohio_2018 cohort may be significantly different between at least two prediction models. In those cases, the post-hoc Nemenyi test was performed for pair-wise comparisons between prediction models. Results of the Nemenyi tests are then visualised using CD diagrams, as shown in Figure 6.1. In each CD diagram, at a

Table 6.7: p-values of the Friedman test for comparing all prediction models for univariate BGL prediction 30 and 60 minutes in advance in Ohio_2018 and Ohio_2020 datasets.

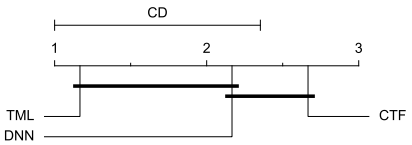
	PH	RMSE	MAE	MCC	SE
Ohio_2018	30 min	1.000	0.607	1.000	0.311
	60 min	**0.006**	**0.030**	0.513	**0.016**
Ohio_2020	30 min	0.223	0.607	0.311	0.607
	60 min	0.311	0.135	0.311	0.069

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

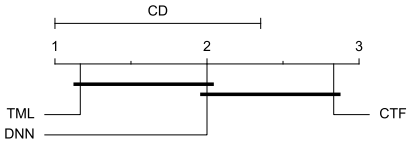
significance level of five percent, prediction models that differ insignificantly are linked by a horizontal line. It can be seen that while the TML model outperformed the CTF model significantly based on their average ranks for the examined metrics, the other pair-wise comparisons were not statistically meaningful.



(a) Metric: RMSE, PH: 60min



(b) Metric: MAE, PH: 60min



(c) Metric: SE, PH: 60min

Figure 6.1: Critical difference diagrams of comparing all prediction models against each other with univariate input over the data contributors of Ohio_2018 dataset based on RMSE (a), MAE(b), and SE (d) metrics for BGL prediction 60 minutes in advance.

Multivariate input Table 6.8 presents p-values of the Friedman test based on each evaluation metric of BGL prediction performance using different forecasting approaches with multivariate input. The test was performed separately for each prediction horizon of 30 and 60 minutes and in each Ohio_2018 and Ohio_2020 cohort. The p-values marked in bold font are considered significant at a significance level of five percent, showing that at least two prediction models may differ in the

BGL prediction performance.

Table 6.8: p-values of the Friedman test for comparing all prediction models for multivariate BGL prediction 30 and 60 minutes in advance in Ohio_2018 and Ohio_2020 datasets.

	PH	RMSE	MAE	MCC	SE
Ohio_2018	30 min	**0.006**	**0.006**	0.115	**0.011**
	60 min	**0.011**	**0.009**	**0.030**	**0.006**
Ohio_2020	30 min	0.223	0.607	0.846	1.000
	60 min	**0.006**	**0.009**	**0.042**	**0.011**

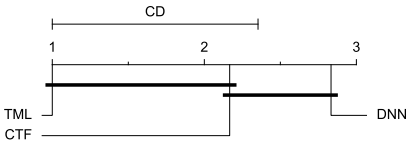
Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Considering the presented results in Table 6.8 and a significance level of five percent, it can be inferred that among different examined cases for comparing prediction approaches regarding evaluation metrics, prediction horizons, and datasets, at least two prediction approaches may perform differently for BGL prediction 60 minutes in advance in both Ohio_2018 and Ohio_2020 datasets based on all the evaluation metrics. Also, there are significant p-values for comparing different prediction models in predicting BGL 30 minutes in advance in Ohio_2018 dataset based on RMSE, MAE, and SE metrics.

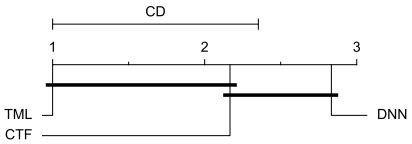
The post-hoc Nemenyi test was conducted for each mentioned case to compare the prediction models in a pair-wise manner. The results of post-hoc tests are graphically presented in CD diagrams, as demonstrated in Figures 6.2 and 6.3 for cohorts Ohio_2018 and Ohio_2020, respectively. A horizontal line connects prediction models that differ insignificantly (with a significance level of five percent).

Figures 6.2a and 6.2b show that the TML model, while performing similarly to the CTF model, outperformed the DNN model significantly for predicting BGL in the Ohio_2018 dataset 30 minutes in advance based on RMSE and MAE metrics, respectively. From Figures 6.2c, 6.2d, and 6.3d it can be seen that the TML model statistically significantly outperformed both CTF and DNN models in the Ohio_2018 dataset based on SE metric for prediction horizon of 30 minutes and based on RMSE for a prediction horizon of 60 minutes, and in Ohio_2020 dataset based on SE metric for a prediction horizon of 60 minutes, respectively. Figures 6.2e, 6.2f, 6.2g, 6.3a, and 6.3b show that while the TML model performed similarly to the DNN model, it outperformed the CTF model significantly for the prediction horizon of 60 minutes in the Ohio_2018 dataset, based on MAE, MCC, and SE

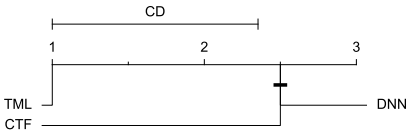
metrics, and also, in the Ohio_2020 dataset, based on RMSE and MAE metrics, respectively. Although based on Table 6.8, the result of the Friedman test calculated based on the MCC metric in the Ohio_2020 dataset for a prediction horizon of 60 minutes was significant, Figure 6.3c shows that for the mentioned case, there was not a significant difference between BGL prediction performance using different prediction models. Also, Table 6.8 and Figures 6.2 and 6.3 reveal that the CTF and DNN models performed similarly for BGL prediction 30 and 60 minutes in advance using multivariate input in both Ohio_2018 and Ohio_2020 cohorts.



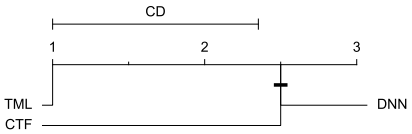
(a) Metric: RMSE, PH: 30min



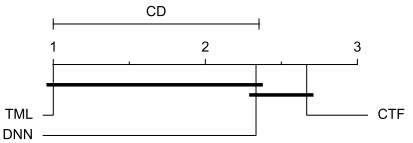
(b) Metric: MAE, PH: 30min



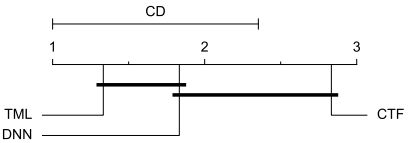
(c) Metric: SE, PH: 30min



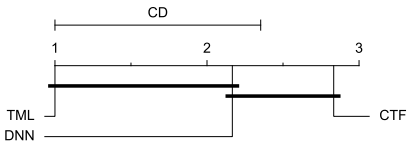
(d) Metric: RMSE, PH: 60min



(e) Metric: MAE, PH: 60min



(f) Metric: MCC, PH: 60min

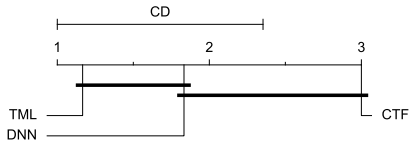


(g) Metric: SE, PH: 60min

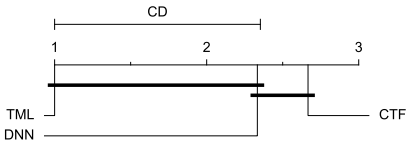
Figure 6.2: Critical difference diagrams of comparing all prediction models against each other with multivariate input over the data providers in Ohio_2018 dataset based on RMSE (a), MAE (b), and SE (c) metrics for BGL prediction 30 minutes in advance and based on RMSE (d), MAE (e), MCC (f), and SE (g) metrics for BGL prediction 60 minutes in advance.

6.3.2.2 Computational cost

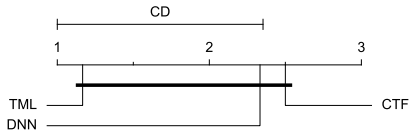
When comparing different prediction models the computational cost of retraining them needs to be considered. The developed models are based on patterns in BGL



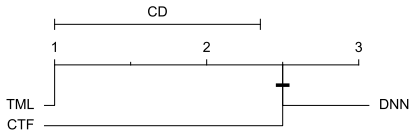
(a) Metric: RMSE, PH: 60min



(b) Metric: MAE, PH: 60min



(c) Metric: MCC, PH: 60min



(d) Metric: SE, PH: 60min

Figure 6.3: Critical difference diagrams of comparing all prediction models against each other with multivariate input over the data providers in Ohio_2020 dataset based on RMSE (a), MAE (b), MCC (c), and SE (d) metrics for BGL prediction 60 minutes in advance.

data. Hence, the models do not have indefinite validity, and readjustments are required following changes in the BGL patterns. The computational costs of different prediction models on a standard laptop computer with a core i7 2.8 GHz processor, an NVIDIA GeForce GTX 1050 Ti GPU, and a 16 GB of RAM were measured. Table 6.9 shows the average execution time for training different models across all six patients in each cohort for each input and prediction horizon. The results illustrate that the TML model is the fastest and the DNN model is the slowest model for retraining purposes.

Table 6.9: The average training time (seconds) for models using different approaches across all patients in each cohort for each input and prediction horizon.

	Model	Univariate		Multivariate	
		PH:30 min	PH:60 min	PH:30 min	PH:60 min
Ohio_2018	CTF	277	289	502	530
	TML	10	11	20	16
	DNN	2057	2094	2051	2100
Ohio_2020	CTF	323	327	558	569
	TML	7	11	14	16
	DNN	1948	2099	1963	2149

Note. PH: Prediction horizon; CTF: Classical time series forecasting; TML: Traditional machine learning; DNN: Deep neural network.

6.3.2.3 Brief findings

Review of the results presented in Sections 6.3.1, 6.3.2.1, and 6.3.2.2 shows that in more than half of the examined cases regarding evaluation metrics, prediction horizons, and datasets, especially using a univariate input, the three models performed comparably in BGL prediction. Among the rest of the cases, the TML model achieved the first rank with a significant superiority over at least one other model. In addition, the TML model was also the fastest model to be trained. The CTF and DNN models performed similarly for BGL prediction in all cases. Overall, the results suggest that the TML model is the superior approach for BGL prediction among the different examined data-driven models.

6.3.3 Comparing models' inputs

In this section, effectiveness of univariate and multivariate inputs are compared using different CTF, TML, and DNN approaches, separately. The outcomes of statistical analyses are given and discussed in the following first section. Furthermore, a discussion about the ease and complexity of different inputs for collection and processing is presented. The results are then summarised to draw conclusions.

6.3.3.1 Statistical result

CTF approach Table 6.10 presents p-values related to the Wilcoxon test, which was performed based on each evaluation metric, prediction horizon, and cohort for examining whether the performance of the CTF model for BGL prediction differs statistically significantly using different inputs. With a significance level of five percent, the test outcomes show that exogenous variables did not affect the performance of BGL prediction using the CTF model 60 minutes in advance in the Ohio_2018 dataset and both at 30 and 60 minutes in advance in the Ohio_2020 dataset based on all evaluation metrics. There is only one statistically significant difference (marked with bold font) between univariate and multivariate inputs using the CTF model, which is related to the RMSE metric for predicting the BGL 30 minutes in advance in the Ohio_2018 dataset.

Considering Tables 6.3, 6.4, and 6.10, it can be concluded that, based on the RMSE metric, the CTF model performed worse with exogenous variables compared to univariate BGL prediction 30 minutes in advance over patients in Ohio_2018 dataset.

Table 6.10: P-values of the Wilcoxon test for comparing univariate and multivariate input for the CTF model for BGL prediction 30 and 60 minutes in advance in Ohio_2018 and Ohio_2020 datasets.

	PH	RMSE	MAE	MCC	SE
Ohio_2018	30 min	**0.031**	0.062	0.438	0.094
	60 min	0.312	0.156	0.225	0.094
Ohio_2020	30 min	0.438	0.844	0.500	1.000
	60 min	0.219	0.562	0.686	0.219

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

TML approach Table 6.11 displays p-values of the Wilcoxon test for examining if univariate or multivariate inputs can make a statistically significant difference in BGL prediction performance by applying the TML model. The test was performed over the data contributors of each cohort and was based on each evaluation metric and for each prediction horizon separately. With a significance level of five percent, the test outcome showed that the TML model predicted BGL significantly differently using different inputs for patients in Ohio_2018 dataset based on the SE metric for both prediction horizons. While the TML model performed similarly using different inputs in Ohio_2020 dataset for both prediction horizons.

Table 6.11: P-values of the Wilcoxon test for comparing univariate and multivariate input for the TML model for BGL prediction 30 and 60 minutes in advance in Ohio_2018 and Ohio_2020 datasets.

	PH	RMSE	MAE	MCC	SE
Ohio_2018	30 min	0.062	0.062	0.156	**0.031**
	60 min	0.062	0.062	0.156	**0.031**
Ohio_2020	30 min	0.438	0.562	0.312	0.844
	60 min	0.062	0.156	0.156	0.094

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

Considering Tables 6.3, 6.4, and 6.11, it can be concluded that the TML model generated better BGL predictions according to SE metric using multivariate input compared to univariate input in Ohio_2018 dataset for both prediction horizons of 30 and 60 minutes.

DNN approach Table 6.12 displays the p-values obtained from the Wilcoxon test, which was performed based on each evaluation metric and for each prediction horizon, over the data contributors of each cohort. The test was conducted to determine whether univariate or multivariate input could make a significant difference in BGL prediction performance by applying the DNN model. The results showed that with a significance level of five percent, there was no statistically significant difference between the performance of the DNN model in predicting BGL using univariate or multivariate input in both datasets and for both prediction horizons, according to all examined evaluation metrics.

Table 6.12: P-values of the Wilcoxon test for comparing univariate and multivariate input of the DNN model for BGL prediction 30 and 60 minutes in advance in Ohio_2018 and Ohio_2020 datasets.

	PH	RMSE	MAE	MCC	SE
Ohio_2018	30 min	0.094	0.094	0.688	0.844
	60 min	0.312	0.562	0.094	0.562
Ohio_2020	30 min	0.844	0.844	0.688	0.688
	60 min	1.000	0.562	1.000	0.688

Note. PH: Prediction horizon; RMSE: Root mean square error; MAE: Mean absolute error; MCC: Matthews correlation coefficient; SE: Surveillance error.

6.3.3.2 Ease of data

Another important factor to be considered for comparing input for the BGL prediction task would be ease of data access. It is essential to consider how convenient data collection and preprocessing would be for each input. Developing a BGL prediction model using only data from a CGM sensor, which is a readily accessible tool for T1DM patients, requires automatic data collection with minimum human intervention and facilitates the practicality of implementation regarding computational complications. In BGL prediction using a univariate input, there would be no need for extra effort and cost to acquire data from several sensors and modalities [72, 73, 56, 75, 76]. Also, multivariate input needs further data preprocessing steps, including data scaling up/down and data alignment. Moreover, according to Table 6.9, BGL prediction using multivariate input needs more computational cost. Overall, univariate input is superior to multivariate input in terms of ease of data collection and processing.

6.3.3.3 Brief findings

According to the results in Sections 6.3.1, 6.3.3.1, and 6.3.3.2 the following can be concluded. There was no conclusive evidence as to whether the use of univariate or multivariate input achieves better BGL prediction performance. With the CTF model, adding exogenous variables could make BGL predictions worse. In contrast, with the TML model, multivariate input may improve BGL prediction, or it may not significantly affect the performance of the DNN model. Also, BGL prediction performance was not significantly impacted by univariate or multivariate input in the Ohio_2020 cohort for the three forecasting models and both prediction horizons. Overall, the results reveal that considering exogenous variables, including carbohydrate, bolus, and activity, despite forcing more effort and cost, does not conclusively make a significant improvement in the performance of BGL prediction. It is important to note that this conclusion is based on the examined naive approaches of including exogenous variables. However, applying advanced data fusion approaches may alter the performance of the models and this conclusion.

6.4 Summary

This work investigated the performance of BGL prediction using different data-driven time series forecasting approaches and different inputs to provide insightful findings in the context of BGL prediction. To do so, three prediction models, including ARIMA, SVR, and LSTM related to the three different time series forecasting approaches were developed. The models were trained with univariate and multivariate inputs. These investigations were performed using two Ohio_2018 and Ohio_2020 cohorts separately. The different cases were evaluated using regression-based and clinical-based metrics followed by rigorous statistical analysis.

The obtained results showed that all three prediction models performed comparably in most cases. In the remaining cases, the TML model, which was also the fastest model to train, performed significantly better than the CTF, the DNN or both especially when using multivariate input. Moreover, comparing different inputs for each prediction model showed that collecting and adding extra variables, including carbohydrates, bolus, and activity, causes additional cost and complexity whilst not improving the BGL prediction significantly. In fact, different time series forecasting approaches perform differently for predicting BGL when dealing with multivariate data. The CTF model may perform worse by adding exogenous variables, the TML model may perform better using multivariate input, and the DNN model performs similarly using univariate or multivariate input.

Chapter 7

Conclusions and future directions

This chapter summarises the research carried out throughout this thesis and presents conclusions. Also, some future directions related to the focus of this thesis are recommended and discussed.

7.1 Summary and conclusions

In this thesis, we developed several novel data-driven methodologies for BGL prediction, aiming to address and fill some of the challenges and gaps in the area. As part of this thesis, advanced architectures leveraging cutting-edge AI techniques, including ensemble learning and causality inference, were developed to improve the performance of BGL predictions. We also developed new PA-informed approaches for BGL prediction by extracting information from PA and incorporating this information in prediction models. Also, to contribute to the fundamental choices of the model structure and input, the performance of different data-driven time series forecasting approaches with different inputs for BGL prediction was investigated. All analyses and methods were developed using the publicly available Ohio T1DM dataset. For evaluation and comparison of developed methods, regression-wised and clinical-wised criteria, along with statistical analyses, were performed. A summary of each performed work is provided below.

The work presented in Chapter 3 proposed advanced architectures to predict BGL in people with T1DM that leverage deep learning and ensemble learning. A vanilla LSTM network, a bidirectional LSTM network, and a linear regression model were used as three base-learners. The meta-learning output fusion strategy was then used to integrate base-learner outputs in three different ways, including univariate time series forecasting, multivariate time series forecasting, and two-dimensional data analysis, named Stacking, Multivariate, and Subsequences ap-

proaches, respectively. The Stacking method involved concatenating the output vectors from the base-learners and feeding them into a linear model serving as the meta-learner. As part of the Multivariate approach, base-learner output vectors served as multivariate input for training a multivariate LSTM model as the meta-learner. The meta-learning thus considered univariate time series forecasting as multivariate forecasting. The Subsequences approach considered base-learner output vectors as different subsequences. This resulted in configuring one-dimensional time series forecasting as two-dimensional data analysis using a ConvLSTM as the meta-learner, offering an overview of the feasibility of meta-learning in changing the dimension. The impact of devised meta-learning strategies on the efficacy of BGL prediction was compared and benchmarked with non-ensemble models. The results showed that the proposed ensemble models statistically significantly outperformed the benchmarked non-ensemble models. Also, the two novel meta-learning approaches performed comparably to the well-effective stacked learning approach.

The work presented in Chapter 4 investigated the deployment of causality inference to improve the performance of BGL prediction in T1DM management. In the first phase of the investigation, the relations between BGL and carbohydrates, bolus, and PA were examined in the causality context. For this purpose, the causal relations were quantified using CCM. Afterwards, ECCM was applied to quantify causality for different lags, thus determining the optimal lag of causality for each variable. Next, two new approaches were proposed for utilising causality information in BGL prediction. In the first approach, causality strengths served as weights for impacting variables. The optimal causal lags and the corresponding causality strengths in the second approach were considered as shifts and weights, respectively, for the variables. The performance of BGL prediction with and without causality deployment was evaluated and compared using regression-wised and clinical-wised evaluation metrics and statistical analyses. Overall, the obtained results showed the effectiveness of developed causality-informed models in BGL prediction.

In Chapter 5 we investigated leveraging PA in BGL prediction by developing different approaches for extracting information from PA data and fusing this information with BGL data in signal, feature, and decision levels to find the optimal approach for deploying PA in BGL prediction models. For the signal-level fusion, different automatically-recorded PA data were fused with BGL data. Also, three feature extraction approaches were used to provide feature-level information for incorporating PA in BGL prediction. Moreover, for decision-level fusion, using ensemble learning, the prediction model was trained using BGL and PA data, as base-learners. Then, their predictions were stacked and used as inputs for the

meta-learner. The developed PA-informed approaches were compared with the no-fusion approach and also with each other. Overall, the results of the evaluation and statistical analyses showed that fusing PA information with BGL can statistically significantly improve the performance of BGL prediction.

In Chapter 6, we compared the performance of different data-driven prediction approaches including CTF, TML, and DNN, as well as different inputs, including univariate and multivariate. To do so, three prediction models, including ARIMA, SVR, and LSTM, were developed, each relating to a forecasting approach. The models were trained with BGL data only, as univariate input. Also, the models' counterparts were developed to cope with multivariate input, including BGL, carbohydrate, bolus, and PA data. The different cases were evaluated using regression-wised and clinical-wised metrics followed by rigorous statistical analyses. The obtained results showed that the SVR model, which was also the fastest model to train, occasionally performed better than the ARIMA, the LSTM, or both, especially when using multivariate input.

It is worth noting that in this thesis, we aimed to carefully determine the effectiveness of utilising advanced AI approaches, including ensemble, causality, and activity fusion. To do so, the approaches developed in Chapters 3, 4, and 5 were rigorously compared with their exact counterparts without using those AI approaches. However, considering the performance of different methods in the second BGL prediction challenge in Table 2.2, shows that our developed methods produced comparable results with the literature.

All in all, the main findings of this thesis can be expressed as follows:

- Ensemble learning, as an advanced AI approach, leveraging different base-learner models can improve the performance of a single BGL prediction model. Hence, it is suggested that by considering adding a bit of complexity, applying ensemble learning can be effective in the BGL prediction task.

- The enhancement of developed causality-based approaches for the BGL prediction task was not considerable. Hence, for BGL prediction tasks, applying developed causality at this stage would not be an efficient way to improve prediction performance. Future research may use these approaches as evidence to leverage this advanced concept to predict BGLs.

- Fusing PA information with BGL can improve the performance of BGL prediction. The results demonstrated the efficacy of leveraging HR data automatically recorded in PA bands to improve BGL prediction performance. Hence,

the BGL prediction task can benefit from using PA bands even by adding the raw HR data to BGL data.

- Regarding the input of BGL prediction models, there is no conclusive decision. Hence, adding exogenous variables, including carbohydrates, bolus, and activity, may not improve the performance of the BGL prediction model unless it is incorporated in an effective way. Therefore, developing effective approaches for leveraging data from other affecting variables is important.

7.2 Future directions

In this thesis, the conducted research contributed to addressing challenges, but there is still much work to be discovered. Future directions related to the focus of this thesis are briefly discussed in the following.

Considering the promising performance that ensemble learning has achieved in the literature and in this thesis, more research into ensemble learning would be beneficial. For example, developing different ensemble learning designs using different base-learners and different fusion output approaches would be useful and lead to enhanced model performance. Also, the performed causality investigation in BGL prediction was a preliminary study and can be developed and improved in different ways. In this regard, further investigation of the deployment of causal inference in BGL prediction by exploring more prediction models and other causality investigation approaches, such as transfer entropy and Peter Clark momentary conditional independence algorithm would be beneficial. Moreover, to obtain more advantages from causal inference, exploring the use of dynamic values of effective variables, instead of their raw data could be considered for future work. Also, causality lags can be informative as prior knowledge and would be considered as a potential tool for improving BGL prediction. Hence, developing and investigating further approaches of leveraging causal lags information in this area would also be beneficial. Moreover, in general, exploring cutting-edge AI strategies for BGL prediction will remain in demand due to AI's fast growth. Hence, other advanced AI strategies such as transfer learning need to be explored in BGL prediction.

Exploring PA, as a challenging factor in T1DM management, and providing insightful findings in the context of BGL prediction is still demanding. The performed PA-related work in this thesis indicated that the intensity categories of PA determined manually were considered informative for BGL prediction models. In light of this, the automation of this procedure by automatically classifying PA data into various intensity levels would be a valuable future direction of research.

Bibliography

[1] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, 2017.

[2] A. D. Association, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 37, no. Supplement_1, pp. S81–S90, 2014.

[3] D. Mellitus, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 28, no. S37, pp. S5–S10, 2005.

[4] I. H. S. Group, "Glucose concentrations of less than 3.0 mmol/l (54 mg/dl) should be reported in clinical trials: a joint position statement of the american diabetes association and the europian association for the study of diabetes," *Diabetologia*, vol. 60, no. 1, pp. 3–6, 2017.

[5] E. Montaser Roushdi Ali *et al.*, "Stochastic seasonal models for glucose prediction in type 1 diabetes," Ph.D. dissertation, 2020.

[6] R. Ajjan, D. Slattery, and E. Wright, "Continuous glucose monitoring: A brief review for primary care practitioners," *Advances in therapy*, vol. 36, no. 3, pp. 579–596, 2019.

[7] J. Beck, D. A. Greenwood, L. Blanton, S. T. Bollinger, M. K. Butcher, J. E. Condon, M. Cypress, P. Faulkner, A. H. Fischl, T. Francis *et al.*, "2017 national standards for diabetes self-management education and support," *The Diabetes Educator*, vol. 46, no. 1, pp. 46–61, 2020.

[8] L. A.-C. Wright and I. B. Hirsch, "Metrics beyond hemoglobin a1c in diabetes management: time in range, hypoglycemia, and other parameters," *Diabetes technology & therapeutics*, vol. 19, no. S2, pp. S–16, 2017.

[9] G. Williams and J. C. Pickup, *Handbook of diabetes*. Wiley-Blackwell, 2004.

[10] B. Balkau, L. Mhamdi, J.-M. Oppert, J. Nolan, A. Golay, F. Porcellati, M. Laakso, E. Ferrannini, and E.-R. S. Group, "Physical activity and insulin sensitivity: the risc study," *Diabetes*, vol. 57, no. 10, pp. 2613–2618, 2008.

[11] H. Tikkanen-Dolenc, J. Wadén, C. Forsblom, V. Harjutsalo, L. M. Thorn, M. Saraheimo, N. Elonen, M. Rosengård-Bärlund, D. Gordin, H. O. Tikkanen *et al.*, "Frequent and intensive physical activity reduces risk of cardiovascular events in type 1 diabetes," *Diabetologia*, vol. 60, no. 3, pp. 574–580, 2017.

[12] 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 management in type 1 diabetes: a consensus statement," *The lancet Diabetes & endocrinology*, vol. 5, no. 5, pp. 377–390, 2017.

[13] W. T. Cefalu and J. L. Leahy, *Insulin therapy*. CRC Press, 2002.

[14] N. Bazaev, A. Pletenev, and K. Pozhar, "Classification of factors affecting blood glucose concentration dynamics," *Biomedical Engineering*, vol. 47, no. 2, p. 100, 2013.

[15] M. Rigla, G. García-Sáez, B. Pons, and M. E. Hernando, "Artificial intelligence methodologies and their application to diabetes," *Journal of diabetes science and technology*, vol. 12, no. 2, pp. 303–310, 2018.

[16] N. J. Nilsson, *Principles of artificial intelligence*. Morgan Kaufmann, 2014.

[17] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: literature review," *Journal of medical Internet research*, vol. 20, no. 5, p. e10775, 2018.

[18] O. Mujahid, I. Contreras, and J. Vehi, "Machine learning techniques for hypoglycemia prediction: trends and challenges," *Sensors*, vol. 21, no. 2, p. 546, 2021.

[19] 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.

[20] J. Brownlee, *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018.

[21] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2194, p. 20200209, 2021.

[22] R. A. Laursen and P. Alo, "Transform diabetes-harnessing transformer-based machine learning and layered ensemble with enhanced training for improved glucose prediction." Master's thesis, University of Agder, 2023.

[23] H. Khadem, M. R. Eissa, H. Nemat, O. Alrezj, and M. Benaissa, "Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy," *Talanta*, vol. 211, p. 120740, 2020.

[24] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "Interpretable machine learning for inpatient covid-19 mortality risk assessments: Diabetes mellitus exclusive interplay," *Sensors*, vol. 22, no. 22, p. 8757, 2022.

[25] H. Khadem, H. Nemat, M. R. Eissa, J. Elliott, and M. Benaissa, "Covid-19 mortality risk assessments for individuals with and without diabetes mellitus: Machine learning models integrated with interpretation framework," *Computers in Biology and Medicine*, vol. 144, p. 105361, 2022.

[26] A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, "Data-driven blood glucose pattern classification and anomalies detection: machine-learning applications in type 1 diabetes," *Journal of medical Internet research*, vol. 21, no. 5, p. e11030, 2019.

[27] S. Ellahham, "Artificial intelligence: the future for diabetes care," *The American journal of medicine*, vol. 133, no. 8, pp. 895–900, 2020.

[28] S. Oviedo Castillo *et al.*, "Forecasting and decision support for type 1 diabetes insulin therapy using machine learning," Ph.D. dissertation, 2019.

[29] J. Vehí, I. Contreras, S. Oviedo, L. Biagi, and A. Bertachi, "Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning," *Health Informatics Journal*, vol. 26, no. 1, pp. 703–718, 2020.

[30] N. S. Tyler, C. Mosquera-Lopez, G. M. Young, J. El Youssef, J. R. Castle, and P. G. Jacobs, "Quantifying the impact of physical activity on future glucose trends using machine learning," *Iscience*, vol. 25, no. 3, p. 103888, 2022.

[31] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[32] M. Wadghiri, A. Idri, T. El Idrissi, and H. Hakkoum, "Ensemble blood glucose prediction in diabetes mellitus: A review," *Computers in Biology and Medicine*, vol. 147, p. 105674, 2022.

[33] P. H. Kassani, L. Xiao, G. Zhang, J. M. Stephen, T. W. Wilson, V. D. Calhoun, and Y. P. Wang, "Causality-based feature fusion for brain neurodevelopmental analysis," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3290–3299, 2020.

[34] D. You, R. Li, S. Liang, M. Sun, X. Ou, F. Yuan, L. Shen, and X. Wu, "Online causal feature selection for streaming features," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[35] P. J. E. A. Javier, M. P. Liponhay, C. V. G. Dajac, and C. P. Monterola, "Causal network inference in a dam system and its implications on feature selection for machine learning forecasting," *Physica A: Statistical Mechanics and its Applications*, vol. 604, p. 127893, 2022.

[36] N. H. Mohammed and T. M. Wolever, "Effect of carbohydrate source on postprandial blood glucose in subjects with type 1 diabetes treated with insulin lispro," *Diabetes research and clinical practice*, vol. 65, no. 1, pp. 29–35, 2004.

[37] J. Xie and Q. Wang, "A data-driven personalized model of glucose dynamics taking account of the effects of physical activity for type 1 diabetes: An in silico study," *Journal of biomechanical engineering*, vol. 141, no. 1, p. 011006, 2019.

[38] M. Zhang, K. B. Flores, and H. T. Tran, "Deep learning and regression approaches to forecasting blood glucose levels for type 1 diabetes," *Biomedical Signal Processing and Control*, vol. 69, p. 102923, 2021.

[39] J. Xie and Q. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for type i diabetes in comparison with classical time-series models," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 11, pp. 3101–3124, 2020.

[40] C. Zecchin, A. Facchinetti, G. Sparacino, and C. Cobelli, "How much is short-term glucose prediction in type 1 diabetes improved by adding insulin delivery and meal content information to cgm data? a proof-of-concept study," *Journal of diabetes science and technology*, vol. 10, no. 5, pp. 1149–1160, 2016.

[41] H. Hameed and S. Kleinberg, "Comparing machine learning techniques for blood glucose forecasting using free-living and patient generated data," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 871–894.

[42] M. Vettoretti, G. Cappon, A. Facchinetti, and G. Sparacino, "Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors," *Sensors*, vol. 20, no. 14, p. 3870, 2020.

[43] S. Oviedo, J. Vehí, R. Calm, and J. Armengol, "A review of personalized blood glucose prediction strategies for T1DM patients," *International journal for numerical methods in biomedical engineering*, vol. 33, no. 6, p. e2833, 2017.

[44] J. I. Hidalgo, J. M. Colmenar, G. Kronberger, S. M. Winkler, O. Garnica, and J. Lanchares, "Data based prediction of blood glucose concentrations using evolutionary methods," *Journal of medical systems*, vol. 41, pp. 1–20, 2017.

[45] V. Felizardo, N. M. Garcia, N. Pombo, and I. Megdiche, "Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review," *Artificial Intelligence in Medicine*, vol. 118, p. 102120, 2021.

[46] C. Novara, N. M. Pour, T. Vincent, and G. Grassi, "A nonlinear blind identification approach to modeling of diabetic patients," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 3, pp. 1092–1100, 2015.

[47] K. Zarkogianni, K. Mitsis, E. Litsa, M.-T. Arredondo, G. Fico, A. Fioravanti, and K. S. Nikita, "Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring," *Medical & biological engineering & computing*, vol. 53, pp. 1333–1343, 2015.

[48] 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.

[49] F. Ståhl and R. Johansson, "Diabetes mellitus modeling and short-term prediction based on blood glucose measurements," *Mathematical biosciences*, vol. 217, no. 2, pp. 101–117, 2009.

[50] A. Gani, A. V. Gribok, S. Rajaraman, W. K. Ward, and J. Reifman, "Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 246–254, 2008.

[51] M. Eren-Oruklu, A. Cinar, L. Quinn, and D. Smith, "Estimation of future glucose concentrations with subject-specific recursive linear models," *Diabetes technology & therapeutics*, vol. 11, no. 4, pp. 243–253, 2009.

[52] J. Yang, L. Li, Y. Shi, and X. Xie, "An arima model with adaptive orders for predicting blood glucose concentrations and hypoglycemia," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1251–1260, 2018.

[53] D. A. Finan, H. Zisser, L. Jovanovic, W. C. Bevier, and D. E. Seborg, "Practical issues in the identification of empirical models from simulated type 1 diabetes data," *Diabetes technology & therapeutics*, vol. 9, no. 5, pp. 438–450, 2007.

[54] M. Eren-Oruklu, A. Cinar, D. K. Rollins, and L. Quinn, "Adaptive system identification for estimating future glucose concentrations and hypoglycemia alarms," *Automatica*, vol. 48, no. 8, pp. 1892–1897, 2012.

[55] F. Prendin, J.-L. Díez, S. Del Favero, G. Sparacino, A. Facchinetti, and J. Bondia, "Assessment of seasonal stochastic local models for glucose prediction without meal size information under free-living conditions," *Sensors*, vol. 22, no. 22, p. 8682, 2022.

[56] J. Martinsson, A. Schliep, B. Eliasson, and O. Mogren, "Blood glucose prediction with variance estimation using recurrent neural networks," *Journal of Healthcare Informatics Research*, vol. 4, no. 1, pp. 1–18, 2020.

[57] 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)*. IEEE, 2017, pp. 2887–2891.

[58] W. Seo, Y.-B. Lee, S. Lee, S.-M. Jin, and S.-M. Park, "A machine-learning approach to predict postprandial hypoglycemia," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–13, 2019.

[59] P. Calhoun, R. A. Levine, and J. Fan, "Repeated measures random forests (rmrf): Identifying factors associated with nocturnal hypoglycemia," *Biometrics*, vol. 77, no. 1, pp. 343–351, 2021.

[60] S. Oviedo, I. Contreras, C. Quirós, M. Giménez, I. Conget, and J. Vehi, "Risk-based postprandial hypoglycemia forecasting using supervised learning," *International journal of medical informatics*, vol. 126, pp. 1–8, 2019.

[61] D. Dave, D. J. DeSalvo, B. Haridas, S. McKay, A. Shenoy, C. J. Koh, M. Lawley, and M. Erraguntla, "Feature-based machine learning model for real-time hypoglycemia prediction," *Journal of Diabetes Science and Technology*, vol. 15, no. 4, pp. 842–855, 2021.

[62] Y. Jin, F. Li, V. G. Vimalananda, H. Yu *et al.*, "Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: empirical study," *JMIR medical informatics*, vol. 7, no. 4, p. e14340, 2019.

[63] K. Plis, R. Bunescu, 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.

[64] M. Cescon, R. Johansson, and E. Renard, "Subspace-based linear multi-step predictors in type 1 diabetes mellitus," *Biomedical Signal Processing and Control*, vol. 22, pp. 99–110, 2015.

[65] A. Bertachi, L. Biagi, I. Contreras, N. Luo, and J. Vehí, "Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks." in *3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ IJCAI-ECAI 2018, 13 July 2018*, 2018, pp. 85–90.

[66] C. Marling and R. C. Bunescu, "The OhioT1DM dataset for blood glucose level prediction," in *3rd International Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2018, pp. 60–63.

[67] C. Marling and R. Bunescu, "The OhioT1DM dataset for blood glucose level prediction: Update 2020," in *5th International Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2020, pp. 71–74.

[68] T. D. R. in Children Network (DirecNet) Study Group. (1999) Diabetes research in children network (direcnet). [Online]. Available: https://public.jaeb.org/direcnet

[69] E. Lehmann, "Preliminary experience with the internet release of aida—an interactive educational diabetes simulator," *Computer methods and programs in biomedicine*, vol. 56, no. 2, pp. 109–132, 1998.

[70] 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.

[71] H. Nemat, H. Khadem, M. R. Eissa, J. Elliott, and M. Benaissa, "Blood glucose level prediction: Advanced deep-ensemble learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2758–2769, 2022.

[72] J. B. Ali, T. Hamdi, N. Fnaiech, V. Di Costanzo, F. Fnaiech, and J.-M. Ginoux, "Continuous blood glucose level prediction of type 1 diabetes based on artificial neural network," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 4, pp. 828–840, 2018.

[73] T. Hamdi, J. B. Ali, V. Di Costanzo, F. Fnaiech, E. Moreau, and J.-M. Ginoux, "Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 2, pp. 362–372, 2018.

[74] F. D'Antoni, M. Merone, V. Piemonte, P. Pozzilli, G. Iannello, and P. Soda, "Early experience in forecasting blood glucose levels using a delayed and auto-regressive jump neural network," in *2019 IEEE 18th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2019, pp. 394–402.

[75] G. Alfian, M. Syafrudin, M. Anshari, F. Benes, F. T. D. Atmaji, I. Fahrurrozi, A. F. Hidayatullah, and J. Rhee, "Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1586–1599, 2020.

[76] H. V. Dudukcu, M. Taskiran, and T. Yildirim, "Blood glucose prediction with deep neural networks using weighted decision level fusion," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 3, pp. 1208–1223, 2021.

[77] 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.

[78] 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.

[79] J. Jeon, P. J. Leimbigler, G. Baruah, M. H. Li, Y. Fossat, and A. J. Whitehead, "Predicting glycaemia in type 1 diabetes patients: experiments in feature engineering and data imputation," *Journal of Healthcare Informatics Research*, vol. 4, no. 1, pp. 71–90, 2020.

[80] W. L. Clarke, "The original clarke error grid analysis (ega)," *Diabetes technology & therapeutics*, vol. 7, no. 5, pp. 776–779, 2005.

[81] 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.

[82] D. C. Klonoff, C. Lias, R. Vigersky, W. Clarke, J. L. Parkes, D. B. Sacks, M. S. Kirkman, B. Kovatchev, and E. G. Panel, "The surveillance error grid," *Journal of Diabetes Science and Technology*, vol. 8, no. 4, pp. 658–672, 2014.

[83] K. Li, J. Daniels, C. Liu, P. Herrero, and P. Georgiou, "Convolutional recurrent neural networks for glucose prediction," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 603–613, 2019.

[84] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Personalized blood glucose prediction for type 1 diabetes using evidential deep learning and meta-learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 193–204, 2022.

[85] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "Blood glucose level time series forecasting: Nested deep ensemble learning lag fusion," *Bioengineering*, vol. 10, no. 4, p. 487, 2023.

[86] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[87] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Causality analysis in type 1 diabetes mellitus with application to blood glucose level prediction," *Computers in Biology and Medicine*, p. 106535, 2023.

[88] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "Interdependent system topologies for deep learning nonlinear time series forecasting," *Neural Networks*.

[89] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[90] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[91] R. Fisher, "Statistical methods and scientific induction," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 17, no. 1, pp. 69–78, 1955.

[92] P. B. Nemenyi, *Distribution-free multiple comparisons.* Princeton University, 1963.

[93] 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, 2020.

[94] 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," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[95] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "Lstm and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 706–712.

[96] J. Daniels, P. Herrero, and P. Georgiou, "A multitask learning approach to personalized blood glucose prediction," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 436–445, 2021.

[97] 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.

[98] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and intelligent laboratory systems*, vol. 39, no. 1, pp. 43–62, 1997.

[99] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[100] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[101] J. Brownlee, *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery, 2016.

[102] A. Krenker, J. Bešter, and A. Kos, "Introduction to the artificial neural networks," *Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech*, pp. 1–18, 2011.

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

[104] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[105] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, pp. 1–39, 2022.

[106] T. Zhu, K. Li, P. Herrero, J. Chen, and P. Georgiou, "A deep learning algorithm for personalized blood glucose prediction." in *3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@ IJCAI-ECAI 2018, 13 July 2018*, 2018, pp. 64–78.

[107] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[108] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[109] S.-M. Lee, D.-Y. Kim, and J. Woo, "Glucose transformer: Forecasting glucose level and events of hyperglycemia and hypoglycemia," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1600–1611, 2023.

[110] A. Bhimireddy, P. Sinha, B. Oluwalade, J. W. Gichoya, and S. Purkayastha, "Blood glucose level prediction as time-series modeling using sequence-to-sequence neural networks." CEUR Workshop Proceedings, 2020.

[111] M. De Bois, M. A. El Yacoubi, and M. Ammi, "Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people," *Computer Methods and Programs in Biomedicine*, vol. 199, p. 105874, 2021.

[112] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[113] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[114] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.

[115] 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, pp. 79–84, 2018.

[116] K. Saiti, M. Macaš, L. Lhotská, K. Štechová, and P. Pithová, "Ensemble methods in combination with compartment models for blood glucose level prediction in type 1 diabetes mellitus," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105628, 2020.

[117] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Data fusion of activity and CGM for predicting blood glucose levels," in *Proceedings of the 5th*

Annual Workshop on Knowledge Discovery in Healthcare Data, vol. 2675, 2020, pp. 120–124.

[118] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "Multi-lag stacking for blood glucose level prediction," in *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2020, pp. 146–150.

[119] S. Langarica, M. Rodriguez-Fernandez, F. Nunez, and F. J. Doyle III, "A meta-learning approach to personalized blood glucose prediction in type 1 diabetes," *Control Engineering Practice*, vol. 135, p. 105498, 2023.

[120] J. Pearl *et al.*, "Models, reasoning and inference," *Cambridge, UK: CambridgeUniversityPress*, vol. 19, p. 2, 2000.

[121] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang, "Causality learning: A new perspective for interpretable machine learning," *arXiv preprint arXiv:2006.16789*, 2020.

[122] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.

[123] M. Eichler, "Causal inference with multiple time series: principles and problems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1997, p. 20110613, 2013.

[124] E. Siggiridou, C. Koutlis, A. Tsimpiris, and D. Kugiumtzis, "Evaluation of granger causality measures for constructing networks from multivariate time series," *Entropy*, vol. 21, no. 11, p. 1080, 2019.

[125] T. Edinburgh, S. J. Eglen, and A. Ercole, "Causality indices for bivariate time series data: a comparative review of performance," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 8, p. 083111, 2021.

[126] H. Chen, B. Y. Chang, M. A. Naiel, G. Younes, S. Wardell, S. Kleinikkink, and J. S. Zelek, "Causal discovery from sparse time-series data using echo state network," *arXiv preprint arXiv:2201.02933*, 2022.

[127] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[128] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[129] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *science*, vol. 338, no. 6106, pp. 496–500, 2012.

[130] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[131] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[132] Y. Hmamouche, P. Przymus, A. Casali, and L. Lakhal, "GFSM: a feature selection method for improving time series forecasting," *International Journal On Advances in Systems and Measurements*, vol. 10, no. 3, pp. 254–264, 2017.

[133] Y. Hmamouche, A. Casali, and L. Lakhal, "A causality based feature selection approach for multivariate time series forecasting," in *DBKDA 2017, The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications*, 2017.

[134] M. He, W. Gu, Y. Kong, L. Zhang, C. J. Spanos, and K. M. Mosalam, "Causalbg: Causal recurrent neural network for the blood glucose inference with iot platform," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 598–610, 2019.

[135] K. Bach, R. Bunescu, C. Marling, and N. Wiratunga, "Preface the 5th international workshop on knowledge discovery in healthcare data (kdh)," in *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2020, pp. 1–4.

[136] J. Freiburghaus, A. Rizzotti-Kaddouri, and F. Albertetti, "A deep learning approach for blood glucose prediction of type 1 diabetes," in *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2020, pp. 131–135.

[137] H. Rubin-Falcone, I. Fox, and J. Wiens, "Deep residual time-series forecasting: Application to blood glucose prediction." *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 20, pp. 105–109, 2020.

[138] H. Hameed and S. Kleinberg, "Investigating potentials and pitfalls of knowledge distillation across datasets for blood glucose forecasting," in *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, 2020.

[139] T. Zhu, X. Yao, K. Li, P. Herrero, and P. Georgiou, "Blood glucose prediction for type 1 diabetes using generative adversarial networks," in *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2020, pp. 90–94.

[140] 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." *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 45, pp. 136–140, 2020.

[141] R. Bevan and F. Coenen, "Experiments in non-personalized future blood glucose level prediction," in *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 2675, 2020, pp. 100–104.

[142] X. Sun, M. M. Rashid, M. Sevil, N. Hobbs, R. Brandt, M.-R. Askari, A. Shahidehpour, and A. Cinar, "Prediction of blood glucose levels for people with type 1 diabetes using latent-variable-based model." *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 20, pp. 115–119, 2020.

[143] J. Daniels, P. Herrero, and P. Georgiou, "Personalised glucose prediction via deep multitask networks." *Proceedings of the 5th Annual Workshop on Knowledge Discovery in Healthcare Data*, vol. 20, pp. 110–114, 2020.

[144] I. Rodríguez-Rodríguez, J.-V. Rodríguez, J.-M. Molina-García-Pardo, M.-Á. Zamora-Izquierdo, M.-T. M.-I. I. Martínez-Inglés *et al.*, "A comparison of different models of glycemia dynamics for improved type 1 diabetes mellitus management with advanced intelligent analysis in an internet of things context," *Applied Sciences*, vol. 10, no. 12, p. 4381, 2020.

[145] S. Cho, E. M. Aiello, B. Ozaslan, M. C. Riddell, P. Calhoun, R. L. Gal, and F. J. Doyle III, "Design of a real-time physical activity detection and

classification framework for individuals with type 1 diabetes," *Journal of Diabetes Science and Technology*, p. 19322968231153896, 2023.

[146] M. Cescon, D. Choudhary, J. E. Pinsker, V. Dadlani, M. M. Church, Y. C. Kudva, F. J. Doyle III, and E. Dassau, "Activity detection and classification from wristband accelerometer data collected on people with type 1 diabetes in free-living conditions," *Computers in biology and medicine*, vol. 135, p. 104633, 2021.

[147] L. Dénes-Fazakas, M. Siket, L. Szilágyi, L. Kovács, and G. Eigner, "Detection of physical activity using machine learning methods based on continuous blood glucose monitoring and heart rate signals," *Sensors*, vol. 22, no. 21, p. 8568, 2022.

[148] B. Ozaslan, S. Patek, and M. Breton, "Quantifying the effect of antecedent physical activity on prandial glucose control in type 1 diabetes: defining exercise on board," in *Proceedings of the Abstracts from ATTD 2017 10th International Conference on Advanced Technologies & Treatments for Diabetes, Paris, France*, 2017, pp. 15–18.

[149] B. Ozaslan, S. D. Patek, and M. D. Breton, "Impact of daily physical activity as measured by commonly available wearables on mealtime glucose control in type 1 diabetes," *Diabetes technology & therapeutics*, vol. 22, no. 10, pp. 742–748, 2020.

[150] B. Ozaslan, S. D. Patek, C. Fabris, and M. D. Breton, "Automatically accounting for physical activity in insulin dosing for type 1 diabetes," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105757, 2020.

[151] N. Hobbs, S. Samadi, M. Rashid, A. Shahidehpour, M. R. Askari, M. Park, L. Quinn, and A. Cinar, "A physical activity-intensity driven glycemic model for type 1 diabetes," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107153, 2022.

[152] 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.

[153] "The 3rd international workshop on knowledge discovery in healthcare data," https://sites.google.com/view/kdhd-2018/bglp-challenge, 2018.

[154] M. F. Rabby, Y. Tu, M. I. Hossen, I. Lee, A. S. Maida, and X. Hei, "Stacked lstm based deep recurrent neural network with kalman smoothing for blood glucose prediction," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–15, 2021.

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

[156] X. Yan and X. Su, *Linear regression analysis: theory and computing*. World Scientific, 2009.

[157] Y. Cao, M. Raoof, S. Montgomery, J. Ottosson, and I. Näslund, "Predicting long-term health-related quality of life after bariatric surgery using a conventional neural network: A study based on the scandinavian obesity surgery registry," *Journal of clinical medicine*, vol. 8, no. 12, p. 2149, 2019.

[158] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara, "Distinguishing time-delayed causal interactions using convergent cross mapping," *Scientific reports*, vol. 5, no. 1, pp. 1–9, 2015.

[159] K. M. Metcalf, A. Singhvi, E. Tsalikian, M. J. Tansey, M. B. Zimmerman, D. W. Esliger, and K. F. Janz, "Effects of moderate-to-vigorous intensity physical activity on overnight and next-day hypoglycemia in active adolescents with type 1 diabetes," *Diabetes Care*, vol. 37, no. 5, pp. 1272–1278, 2014.

[160] J.-W. van Dijk, T. M. Eijsvogels, J. Nyakayiru, T. H. Schreuder, M. T. Hopman, D. H. Thijssen, and L. J. van Loon, "Glycemic control during consecutive days with prolonged walking exercise in individuals with type 1 diabetes mellitus," *Diabetes research and clinical practice*, vol. 117, pp. 74–81, 2016.

[161] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020.

[162] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy," *Talanta*, vol. 243, p. 123379, 2022.

[163] M. C. Riddell, D. P. Zaharieva, L. Yavelberg, A. Cinar, and V. K. Jamnik, "Exercise and the development of the artificial pancreas: one of the more difficult series of hurdles," *Journal of diabetes science and technology*, vol. 9, no. 6, pp. 1217–1226, 2015.

[164] T. Zhu, C. Uduku, K. Li, P. Herrero, N. Oliver, and P. Georgiou, "Enhancing self-management in type 1 diabetes with wearables and deep learning," *npj Digital Medicine*, vol. 5, no. 1, p. 78, 2022.

[165] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

[166] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of econometrics*, vol. 54, no. 1-3, pp. 159–178, 1992.

[167] I. Rodríguez-Rodríguez, J.-V. Rodríguez, W. L. Woo, B. Wei, and D.-J. Pardo-Quiles, "A comparison of feature selection and forecasting machine learning algorithms for predicting glycaemia in type 1 diabetes mellitus," *Applied Sciences*, vol. 11, no. 4, p. 1742, 2021.

[168] M. P. Reymann, E. Dorschky, B. H. Groh, C. Martindale, P. Blank, and B. M. Eskofier, "Blood glucose level prediction based on support vector regression using mobile platforms," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 2990–2993.

[169] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.