UiT The Arctic University of Norway

Faculty of Science and Technology
Department of Computer Science

Improving Blood Glucose Prediction for People with T1DM During Physical Activity Using Machine Learning on Participant Collected Data

Doyoung Oh

INF-3990-1 Master's Thesis in Computer Science - May 2024

Supervisors
Main supervisor: Professor. Eirik Årsand [UiT/NSE]
Co-supervisor: Associate Professor. André Henriksen [UiT]
Co-supervisor: PhD candidate. Miriam Wolff [NTNU]
Co-supervisor: Dr. Phuong Dinh Ngo [NSE]

I would like to express my sincere gratitude and appreciation to my thesis supervisors, Eirik Årsand, André Henriksen, Miriam Kopperstad Wolff, and Phuong Dinh Ngo. They were incredibly helpful and encouraging throughout the entire journey. I am thankful for their feedback, suggestions, and support.

Abstract

For people with Type 1 Diabetes Mellitus (T1DM), engaging in physical activities (PA) presents unique challenges. The aim of this thesis was to improve the prediction of blood glucose (BG) levels for individuals with T1DM during and after PA. The study began with a literature review to guide the research direction and understand existing prediction models. Then particular emphasis was placed on analyzing papers that provided open-source code, allowing validation of these models using the OhioT1DM dataset and data collected from participants. The GluPredKit platform, an open-source blood glucose prediction framework, was used to streamline the process of data handling, training, and evaluating BG prediction models in Python. The study progressed by training and evaluating various machine learning (ML) models with data from two participants with T1DM. Finally, Physiological Hybrid models and various ensemble models were implemented to observe performance improvement during physical activities.

The Physiological Hybrid model did not improve the predictions during PA compared to the conventional ML models. Although ensemble modeling provided a slight improvement in prediction performance, no ensemble consistently outperformed others, indicating a need for further refinement. Additionally, traditional metrics like Root Mean Squared Error (RMSE) were found to be insufficient in accurately assessing model performance during PA. This prompted the introduction of an additional evaluation method, trajectory plots.

Despite these advancements, this study has several limitations, including the small sample size and heavy reliance on data from smartwatches. As a result, future research should focus on recruiting more participants, refining metrics to better assess ML model performance during PA, and exploring innovative modeling approaches to achieve improved outcomes.

Contents

List of Figures

List of Tables

List of Listings

Acronyms

Acronym	Meaning
AUC	Area under the curve
BG	Blood Glucose
CGM	Continuous glucose monitoring
CRNN	Convolutional recurrent neural network
MDI	Multiple daily injection
ML	Machine Learning
MLP	Multilayer Perceptron
PA	Physical Activity
PH	Prediction Horizon
PLSR	Partial least squares regression
RMSE	Root Mean Squared Error
T1DM	Type 1 diabetes mellitus

1 Introduction

1.1 Background

Overview of existing knowledge about physical activities in T1DM management

For people with T1DM, there are challenges to physical activity (PA). For example, adults with diabetes are at an increased risk for underlying heart disease, and exercise can potentially trigger angina attacks in individuals with preexisting cardiovascular conditions [1].

For people who have had diabetes for a long time and have developed certain complications, exercise must be approached with caution due to the potential risk of hypoglycemia. For example, intense physical activity in individuals with proliferative retinopathy can increase the risk of complications, such as retinal detachment or retinal and vitreous hemorrhages [2].

Why do people with T1DM still need to exercise?

Despite the potential adverse effects of exercise, exercise is beneficial for people with type 1 diabetes (T1DM) for several reasons.

Regular physical activity can improve insulin sensitivity, making your body's cells more receptive to insulin. The result can be better control of blood glucose levels and a reduction in the need for insulin by reducing the risk of insulin resistance [3].

Exercise has cardiovascular benefits for everyone, but it's especially important for people with T1DM who may be at higher risk of cardiovascular complications. Livingstone et al. [4] found that T1DM continues to be associated with higher cardiovascular disease and mortality rates than the non-diabetic population. Regular physical activity can improve the health of the heart, lower blood pressure, and reduce the risk of heart disease.

Weight-bearing exercise, such as walking or resistance training, can improve bone density and reduce the risk of osteoporosis, which affects some people with T1DM. For example, adolescents with T1DM may not reach their potential peak bone mass, putting them at greater risk of fractures [5]. Adult men with T1DM have reduced bone density at the hip, femoral neck, and spine compared with age-matched controls [6].

The challenge of PA in T1DM

Physical activity can be beneficial for people with type 1 diabetes, but it comes with certain challenges and risks. The level of difficulty depends on various factors, including the individual's overall health and the type of physical activity they engage in. The four main challenges are acute hypoglycemia, post-exercise hypoglycemia, hyperglycemia, and exercise-induced ketosis.

- **Acute Hypoglycemia**
  Exercise can have a significant impact on blood glucose levels in people with diabetes, potentially leading to acute hypoglycemia. Physical activity naturally increases the body's demand for energy, prompting muscles to use more glucose, which can result in a drop in blood glucose levels.

  Moreover, exercise can enhance insulin sensitivity, meaning that cells respond more readily to insulin. This heightened sensitivity increases glucose uptake by muscles, potentially causing hypoglycemia if insulin doses are not adjusted accordingly.

  Additionally, the timing of exercise in relation to meals and medication must be carefully managed. Exercising on an empty stomach or soon after taking insulin can elevate the risk of hypoglycemia.

- **Post-exercise Hypoglycemia**
  Post-exercise hypoglycemia, also known as exercise-induced hypoglycemia, occurs when blood glucose levels drop to abnormally low levels after physical activity. This condition can be particularly concerning for people with diabetes, especially those on insulin or other blood glucose-lowering medications. During exercise, the muscles use glucose for energy, which, especially during intense or prolonged workouts, can lead to a significant drop in blood glucose levels.

Additionally, after exercise, peripheral muscles increase their uptake of glucose as they rebuild their glycogen stores, potentially causing hypoglycemia for up to 24 hours post-exercise.

Moreover, short-acting insulin, designed for rapid action and with a relatively short duration, can be absorbed at a higher rate when injected into limbs involved in physical activity, further contributing to the risk of exercise-induced hypoglycemia.

- **Acute hyperglycemia**
  Conversely, certain factors can lead to elevated blood glucose levels during or after exercise. These factors include stress hormones, a counterregulatory response, insulin resistance, and the timing of exercise.

  Intense or strenuous exercise can trigger the release of stress hormones such as cortisol and adrenaline. These hormones can prompt the liver to release stored glucose into the bloodstream, potentially causing acute hyperglycemia [7].

  The body's counterregulatory response, which is intended to prevent hypoglycemia, can sometimes cause hyperglycemia. In individuals with diabetes, this response may be exaggerated, leading the liver to release more glucose than needed.

  During exercise, the body's cells can also become temporarily resistant to insulin, hindering glucose uptake into the cells and resulting in elevated blood glucose levels.

  Finally, the timing of exercise relative to meals and medication can impact blood glucose levels. For instance, exercising shortly after a large carbohydrate-rich meal can cause a spike in blood glucose.

- **Exercise-induced ketosis**
  Exercise-induced ketosis in diabetes can be a concern if it becomes excessive or prolonged. While ketosis itself is a natural metabolic process that can occur when the body switches to using fat for energy rather than glucose, it can pose potential risks and complications in diabetes, particularly if not managed.

  Diabetic ketoacidosis (DKA) is a serious and potentially life-threatening complication that can occur when ketosis becomes excessive and uncontrolled. It is more common in people with type 1 diabetes, but can also occur in people with type 2 diabetes [8]. DKA occurs when there is a severe lack of insulin, leading to a build-up of ketones in the blood. High ketone levels can make the blood acidic, causing a range of symptoms including nausea, vomiting, abdominal pain, confusion, and even unconsciousness.

Physical activity has complex effects on the body, so it's crucial to take into account a variety of factors that can influence blood glucose levels before, during, and after exercise. The

type of exercise (such as low-intensity versus moderate- to high-intensity), the duration of the workout, the patient's level of physical conditioning, their prior diet, and the degree of insulin deficiency all play significant roles.

1.2 Scope and research problem

In this project, I aim to develop a predictive machine-learning model to predict the blood glucose level of people with Type 1 Diabetes Mellitus (T1DM). The prediction algorithm should be developed to improve the accuracy and thus to improve the self-management of people with T1DM during physical activity. This project focuses on incorporating physical activity into prediction models for blood glucose levels in people with Type 1 Diabetes Mellitus.

The scope of this project is divided into two parts. The first part involves comparing models from literature review papers that have publicly available open-source code. All the selected papers used the same OhioT1DM dataset, which contains 8 weeks of data from 12 individuals with Type 1 diabetes. Each of these individuals was using insulin pump therapy with a continuous glucose monitor (CGM). The dataset provides a variety of data types, including CGM blood glucose readings every 5 minutes, blood glucose levels from periodic self-monitoring (finger sticks), insulin dosages (both bolus and basal), self-reported meal times with carbohydrate estimates, self-reported times of exercise, sleep, work, stress, illness, and physiological data from fitness bands [9].

The second part of the project involves creating and training a predictive model using data from people with Type 1 Diabetes Mellitus (T1DM). This data includes continuous glucose monitor (CGM) readings, insulin injections, carbohydrate intake, and physical activity records. The objective is to develop a model that can accurately predict future blood glucose levels during the PA. To achieve this, firstly, several types of machine learning models were compared to identify which performed best in forecasting blood glucose levels. Then drawing from insights gained in both parts of the project, new strategies were adopted to develop a machine-learning model optimized for performance during physical activity.

Thus this thesis aims to address the following problems.

- Main research problem
  What machine learning techniques can be used to generate accurate predictive models that anticipate how an individual's blood glucose levels may fluctuate during exercise?

Generating accurate predictive models for anticipating an individual's blood glucose levels during exercise involves handling time-series data, understanding complex physiological interactions, and capturing personalized responses. Several machine-learning techniques will be compared for this purpose. It's crucial to note that generating accurate predictive models for blood glucose variability is a challenging task due to individual variability and the complex interplay of physiological factors. Data pre-processing, feature engineering, model selection, and hyper-parameter tuning are critical steps in the development of effective predictive models.

- Sub-problem
  How to measure the performance of a machine learning model during the physical activity Measuring the performance of a machine learning (ML) model during physical activity (PA) involves multiple factors. Traditional metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) can be useful, but they might not provide a complete understanding of performance due to the dynamic fluctuations in blood glucose levels during exercise. To gain a more accurate assessment of model performance, it's essential to include time-dependent metrics and tools that can visualize prediction accuracy over time.

  Moreover, the evaluation should consider the significant variability in blood glucose levels that occurs during physical activity. A broader view will allow for a more thorough assessment of how well the ML model adapts to the rapid changes that can happen during exercise.

2

Literature Review

2.1 Introduction

In this section, I detailed the literature review I conducted to analyze existing scholarly work related to my specific research topic. My research topic encompasses multiple domains, including diabetes management, physical activity, and machine learning. Physical activity is highly related to hypoglycemia events in people with T1DM. The data collected by Zaitcev et al. [10] has characterized the incidence distribution of the different causes of hypoglycemia, with more than 45% of the cases collected being related to physical activity. A literature review has guided me in integrating knowledge from these fields. It served various purposes in the context of my master's thesis, helping me understand the current state of knowledge in the research area. It also allowed me to identify what has already been studied and the challenges that still exist by reviewing existing methods and technologies proposed in the literature. Through this process, I was able to identify and define the key theories and theoretical foundations for my thesis. I also aim to justify my research by showing the need for my study and its contributions to the field.

2.2 Methods

The following are how I conducted my literature review, detailing how I searched the relevant research papers, how I filtered them for inclusion or exclusion, and how I reviewed them.

2.2.1 Search strategy

I searched PubMed, ACM, IEEExplorer, and SCOPUS databases for relevant literature from 1 January 2013 to 31 August 2023. The search keywords were "Type 1 Diabetes", "Physical Activity", and "Prediction". The search query used in IEEExplorer and ACM is:

(“T1DM” OR “T1D” OR "Type 1 Diabetes") AND ("Physical Activity" OR "Exercise") AND (“Prediction” OR "Machine Learning" OR "Predictive Model" OR “Forecast*” OR “Neural Network*” OR “Deep Learning”)

The following query syntax was used for PubMed:

(T1DM[Title/Abstract] OR T1D[Title/Abstract OR Type 1 Diabetes[Title/Abstract]) AND (Physical Activity[Title/Abstract] OR Exercise[Title/Abstract]) AND (Prediction[Title/Abstract] OR Machine Learning[Title/Abstract] OR Predictive Model[Title/Abstract] OR Forecast[Title/Abstract] OR Neural Network[Title/Abstract] OR Deep Learning[Title/Abstract])

For SCOPUS, a different query is used as it does not allow the wildcard character ‘*’. Plus, due to the limitations in using a maximum of 8 boolean connectors per field, the query needed to be divided as below:

TITLE-ABS(“T1DM” OR “T1D” OR "Type 1 Diabetes") AND TITLE-ABS("Physical Activity" OR "Exercise") AND TITLE-ABS(“Prediction” OR "Machine Learning" OR "Predictive Model" OR “Forecast”)
TITLE-ABS(“T1DM” OR “T1D” OR "Type 1 Diabetes") AND TITLE-ABS("Physical Activity" OR "Exercise") AND TITLE-ABS(“Prediction” OR "Deep Learning" OR "Predictive Model" OR “Neural Network”)

A total of 180 records were found, with 37 of them being relevant to this review, and added 7 papers from Knowledge Discovery in Healthcare Data 2020 (KDH 2020), which resulted in a total of 44 studies. The selection process of identification, screening, checking for eligibility, and including additional discoveries is described in Figure 2.1 below.

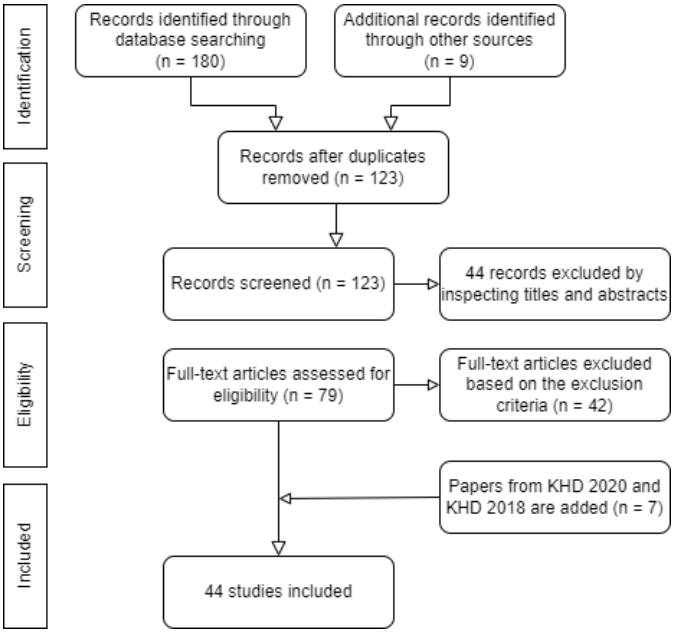

Figure 2.1: The study selection process flow diagram

2.2.2 Inclusion and Exclusion Criteria

In order to include relevant records and to increase the reliability of this review, I included the records that met the following conditions:

- The study population was patients with type 1 diabetes
- The prediction model that included physical activity data as input for training the model
- The main outcomes of the records were detailed and presented with algorithms, or models related to the prediction.

On the other hand, the exclusion criteria are as follows:

- The research topic was an algorithm proposal or improvement of hypoglycemia warning in AP or CGM products
- Abstracts/short papers
- Reviews

2.2.3 Data extraction and Quality assessment

Data were extracted by carefully scheming from the full text of the records that were identified as relevant after the filtering process. For each study, the following data were commonly extracted: first author, year of publication, sample size, input data, prediction horizon(PH), the algorithms or models used, and validation approach.

2.3 Results

Of the 180 records obtained after the initial search, 123 records remained after the removal of duplicate results, 79 records remained after the primary screening of titles and abstracts and were assessed for eligibility, and 42 records were excluded after full-text review. Finally, 7 additional studies from KDH 2020 were included in the review. Following is the summary of the studies based on the extracted data:

References	Publication year	Participant size	PA input	Challenges / Aims	Proposal of study to address the challenge	Modeling approaches	Prediction horizon	Evaluation Metric	Open source code
Afentakis et al. [11]	2023	37	Administered PA (total steps count, active distance, active minutes, total estimated energy expenditure, resting heart rate)	In previous studies, participants were using CGM for glucose monitoring; however, only in one study, they were using MDI regime. In addition, no external validation was conducted in any of the studies apart from a work where two validation data sets were employed.	Used two distinct data sets from participants using CGM and MDI to develop and externally validate ML models.	Random Forests (RF) and Support Vector Machines (SVMs)	NA	ROC-AUC, Sensitivity, Specificity	N
Balakrishnan et al. [12]	2012	NA	Rate of perceived exertion (RPE) values	Recent studies have not incorporated any exercise effects into the personalized models, and these models have been developed using the virtual data only.	Time series BG models with inputs related to exercise (quantified by 0 to 10 revised scale RPE values), meal, and insulin (basal and bolus) were utilized for developing personalized BG models.	Autoregressive moving average exogenous input (ARMAX), ARX and HW models	NA	Cross validation percentage fitness values (% FCVal)	N
Bergford et al. [13]	2023	459	Lower glucose values at the start of exercise (i.e., values < 125 mg/dL) and a greater negative glucose rate of change at the start of exercise (i.e., < -0.5 mg/dL per min), type of PA	Factors for predicting exercise-associated hypoglycemia during exercise is unclear.	Two models were trained to predict hypoglycemia during exercise: a repeated measures random forest (RMRF) model and a repeated measures logistic regression (RMLR) model.	RMLR	NA	AUC, and Brier score were calculated for the training and test data. Balanced accuracy was also determined for the test data when classifying exercises as high or low risk using Youden's index	N
Bertachi et al. [14]	2018	6	Action on board (AOB) using steps	So far there is not any commercial fully-automated system that completely withdraw the burden from patients of taking daily decisions regarding diabetes management	Two different tools that may be used by subjects to support daily decisions regarding diabetes management using ANN and physiological models: i) a tool to provide prediction of BG levels continuously and ii) a tool to predict the occurrence of nocturnal hypoglycemic events.	ANN	30 and 60 min	RMSE	N

Bertachi et al. [15]	2020	10	two features related to physical activity (AOB and estimation of calories burned)	T1D patients using MDI therapy are more exposed to NH than SAP users, thus more effort should be directed towards this group.	Personalized predictive models were generated using two supervised ML algorithms that have been widely applied in supervised learning: multilayer perceptron networks (MLP) and support vector machines (SVM).	MLP and SVM	NA	Averaged sensitivity (SN), specificity (SP), accuracy, and Gmean	N
Bogue-Jimenezn et al. [16]	2022	12	heart rate	Sensors, measuring the dielectric properties of the skin using resonant methods, are under further development to improve sensitivity. There is potential promise in integrating active circuitry and machine learning techniques.	investigate a synergistic approach to accurately predict BGLs by combining noninvasive biometrics measurements with machine learning algorithms	linear regression (LR), Support Vector Regression (SVR), K Nearest Neighbors Regression (KNN), Decision Trees Regression (DTR), bagging trees regression (BTR), Random Forest Regression (RFR), Gaussian process regression (GPR), and Multi-layer Perceptron Regression (MLP).	NA	Clarke Error Grids, RMSE, R2 values	N
Calhound et al. [17]	2020	127	The level of exercise intensity for that day	Partitioning the data by the signs of their average residuals is potentially ineffective	Developed an RF algorithm that can split into all types of predictors and handle repeated measurements	RMRF	NA	Robust Wald statistic and Gini gain	N
Canete et al. [18]	2012	20	The amount of exercise taken was also computed assuming a three-level code: low (50–150 cal/30 min), moderate (150–200 cal/30 min), and strong (>200 cal/30 min).	The wide variability in glucose metabolism between subjects, together with its nonlinear nature, makes it difficult to obtain both a reliable glucose-insulin model for representing individual behavior and its specific controller.	Incorporated artificial neural networks into the control structure	ANN	NA	RMSE	N
Cappon et al. [19]	2020	6	Self-reported physical exercise	Recurrent neural networks such as LSTMs are known to achieve good performance for the specific task of BG prediction, but they lack interpretability.	Exploited SHapley Additive exPlanations (SHAP), i.e., a newly developed approach to interpreting deep learning model predictions	A bidirectional LSTM	30 and 60 min	RMSE, MAE, TG	Y
Cescon et al. [20]	2011	NA	Increased heart rate, respiration rate, and body movements	Blood glucose dynamics is that it heavily varies over time, often quickly and unexpectedly. As a consequence, a linear-time-invariant model may not be sufficient to produce accurate forecasts of future glycemia	Presented online data-driven multi-step ahead predictions of T1DM patient's blood glucose levels, exploiting meal information, insulin dosing, and vital signs	Online subspace-based multi-step predictors	30 min	Variance Accounted For	N

Contreras et al. [21]	2018	6	AOB using steps	the wide range of variability in the glucose dynamics of T1D patients makes the generation of predictive models a challenging and crucial task	a prediction tool based on the grammatical evolution method which introduces multiple features with the aim of dealing with unforeseen changes	Grammatical Evolution Approach	30, 60, and 90 min	RMSE, gRMSE and Clarke error grid zones	N
Daniels et al. [22]	2020	6	Self-reported physical exercise	Typically Deep Learning models require relatively large amounts of data to converge on an appropriate model.	Employ a multitask learning approach in order to improve the performance of the glucose forecasting in a neural network, where each individual is viewed as a task, using shared layers to enable learning from other individuals	MTL approach, CRNN Model	30 and 60 min	RMSE, MAE	Y
De Paoli et al. [23]	2021	6	The days and times in which PA was performed and its type	Despite the excellent performance obtained by predictive models, the prediction of abrupt changes in blood glucose values produced during sports.	The application of a Jump Neural Network to perform a regression task with a univariate approach, in order to reduce, the burden of the patient without requiring them to supply data manually or to wear unnecessary sensors.	Jump Neural Network	30 min	RMSE	N
Ewings et al. [24]	2014	50	METs15 (metabolic equivalent of tasks, a measure of energy expenditure).	The effect of physical activity on internal physiological processes is rarely measurable	A physiologically based model of blood glucose dynamics is developed	Bayesian Network	NA	MCMC	N
Faccioli et al. [25]	2018	6	PA using step counts	A fixed control algorithm, designed on an average patient, could not guarantee satisfactory glycemic control for all possible T1D patients	Compared Multi Input Single Output (MISO) black-box models identified only using meal and insulin information with those identified using also physical activity information	Black-box model	5, 15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165, 180 min	RMSE and Coefficient of Determination (COD)	N
Georga et al. [26]	2012	NA	cumulative energy expenditure (SEE)	The inherent nonlinearity and nonstationarity of the glucose regulatory system limits the predictive capacity (up to 30 min) of the autoregressive models	Random Forests (RF) regression technique is employed to deal with the problem of s.c. glucose prediction in type 1 diabetes based on a multivariate dataset acquired under free-living conditions.	Random Forests (RF) regression model	15, 30, 60 and 120 min	average RMSE, Clarke's Error Grid Analysis (EGA)	N
Georga et al. [27]	2012	27	EE. Wearable body monitoring systems acquire body physiological signals from multiple sensors. This system reports the energy expenditure of daily physical activities or exercise events every 1 min	The intrinsic nonlinearity and nonstationarity of the glucose regulatory system	Systematic work that examines the effect of a number of factors on s.c. glucose prediction in type 1 diabetic patients with the aid of the SVR technique	SVR	15, 30, 60, and 120 min	RMSE	N

Georga et al. [28]	2015	15	Calculated energy expenditure cumulatively every 10 minutes over the last 3 hours.	The existent inter- and intra-patient variability in type 1 diabetes implies the individualization of the predictive models and their continuous adaptation to both biological and environmental changes as well	Examined the concurrent and cumulative impact of the most important predictors of the short-term daily glucose dynamics in a type 1 diabetic individual with the aid of feature ranking. The input is not predefined, but it is selected separately for each patient from a high-dimensional feature set which may result in much simpler models	Random forests (RF) and RReliefF algorithms to rank the candidate feature set. Then, a forward selection procedure to build a glucose predictive model, where features are sequentially added to it in decreasing order of importance. Predictions are performed using support vector regression or Gaussian processes.	30, and 60 min	RMSE	N
Georga et al. [29]	2015	15	Energy expenditure calculated cumulatively every 10 min over the last three hours	Elaborate optimization approaches (e.g. back-propagation, quadratic programming) limit their applicability to glucose prediction where multiple days of patient monitoring are needed to obtain a reliable predictive model. Moreover, such models do not necessarily outperform simple time series models with exogenous inputs.	An extended Kalman filter was proposed to recursively estimate the time-varying coefficients of a patient-specific 3-variate (i.e. glucose level, insulin dose, and meal intake) state-space model.	Extreme learning machine (ELM) and Online sequential ELM are tested with sigmoid activation functions, whereas Kernel ELM and KOS-ELM are tested with a Gaussian kernel	30 min	RMSE, TG, ESOD	N
Georga et al. [30]	2019	NA	The energy expenditure calculated cumulatively every 10 min over the last 3 hours	Nonlinear multivariate dynamical modeling of blood glucose is essential to representing the intrinsic nonlinearity and non-stationarity of the glucose system.	Presented a recursive multivariable kernel adaptive filtering (KAF) approach to personalized short-term glucose prediction in type 1 diabetes	Kernel adaptive filters	5, 15, 30, 45, and 60 min	RMSE and MAPE	N
Jaloli et al. [31]	2022	6	accelerometer (ACC) and electrodermal activity (EDA) signals collected by a wearable device, Physical Activity Intensity	Characterize the effect of physical activity and stress on blood glucose fluctuations by using data collected with wearable devices, and incorporate it into BG predictive models to achieve more accurate BG predictions	Propose a 2-steps approach: in the first step, biomarkers for PA and stress are derived from raw accelerometer and electrodermal activity signals collected by a wearable device. In the second step, we combine the obtained biomarkers with the CGM, meal, and insulin intakes in a multivariate dataset and feed it to our DL-based glucose predictive model to forecast the future BG values	LSTM and CNN-LSTM model	30, 60, and 90 min	RMSE, MAE, R2	N

Jeon et el. [32]	2019	6	Heart rate, steps taken, galvanic skin response, skin temperature, exercise intensity	Unexpected malfunction of monitoring devices and unreliable self-reporting causes data gaps, thus limiting the accuracy of predicting future BG levels.	Explored various imputation methods on the training set of each patient in the OhioT1DM cohort, and compared the prediction accuracy on the test set for each imputation method.	XGBoost model	NA	RMSE, MAE, and Pearson's correlation coefficient (PCC)	N
Khadem et al. [33]	2023	6	Automatically collected PA data using physiological sensors	A quandary is to select the appropriate length of history to be investigated	A compound lag fusion approach by exploiting the potential of nested ensemble learning over typical ensemble learning analysis	Non-stacking, stacking, and nested stacking models	30, and 60 min	MAE, RMSE, MAPE	N
Liu et al. [34]	2018	20	Physical exercise produces significant changes in insulin sensitivity. Since the effect of physical exercise on glucose uptake and insulin sensitivity is not explicitly modeled within the employed minimal model, its effect is taken into account by modifying the sensitivity parameter	Additional information such as meal absorption and physical exercise information can potentially further improve accuracy	Model-based glucose prediction algorithm which uses deconvolution of the CGM signal to estimate some model states in order to improve prediction accuracy. In addition to using CGM data, insulin boluses, and carbohydrate intake information, information about meal absorption and physical exercise is taken into account to further enhance prediction accuracy	LVX algorithm with Kalman filter technique	NA	AUC, RMSE	N
Martínez-Delgado et al. [35]	2023	9	The standard deviation of the acceleration data was used to consider the physical activity of the patient	A proper combination of machine learning models and theoretical physiological absorption models for insulin and carbohydrates could improve convergence and results for deep learning models	Used a recurrent neural network (RNN) based on LSTM cells in order to estimate future levels of blood glucose based on past readings coming from a continuous blood glucose monitor (CGM), insulin injections, and carbohydrate intake and works on different absorption models to process the data available from real patients.	LSTM Based RNN Proposed Architecture	30 and 60 min	RMSE	N
Mirshekarian et al. [36]	2019	NA	HR	Accurate forecasting of blood glucose levels would enable people with T1D to proactively intervene to prevent these conditions from occurring	Improve blood glucose level prediction using recurrent neural networks, using both simulated patient data and data collected from people with T1D on insulin pump therapy with CGM	LSTM	30 and 60 min	RMSE	Y

Mirshekarian et al. [37]	2017	NA	Exercise	The impact that the sensor measurements have on BG prediction performance will depend on the proper modeling of their relations with the other variables in the system. However, reengineering could be very time-consuming and cognitively demanding, while lacking in scalability	It proposed to leverage recent advances in unsupervised feature learning and deep learning in order to build a platform that can seamlessly incorporate any number of physiological variables	RNN approach that uses LSTM units	60 min	RMSE	N
Mosquera-Lopez et al. [38]	2023	50	Self-reported information about the type, timing, duration, and intensity of PA	The exercise in the previous study was performed under highly controlled clinical settings, the accuracy of the developed algorithms may not be as high under real-world conditions.	Quantified the impact of key factors explaining hypoglycemia risk using explainable machine learning models.	Mixed-effects logistic regression, Mixed-effects random forest	15, 30, and 60 min	AUROC	N
Nemat et al. [39]	2020	6	Self-reported physical exercise	Data fusion of activity and CGM data normally results in models with a performance not comparable with those using CGM alone.	Proposed two novel CGM and activity data fusion methods to generate BGL prediction models with performance comparable with those using CGM data alone	Three base regressions(MLP, LTSM, PSLR) and a stacked regression technique	30 and 60 min	RMSE, MAE	Y
Parcerisas et al. [40]	2022	10	The accumulated effects of PA at bedtime and steps	Reducing the occurrence of nocturnal hypoglycaemias (NH)	The algorithms for predicting NH have been optimized for a reduced number of features, using only information from CGM and MDI therapy, thus simplifying the overall system	SVM	NA	MCC, SE, SP, F1score and Gmean, ROC curve	N
Pavan et al. [41]	2020	6	Self-reported physical exercise and acceleration data but only the features with the highest ranks are used to train the model	Over the last two decades, several non-linear algorithms have been tested in this framework, none of these models has stood out from the others in terms of prediction accuracy.	Investigated the impact of hyperparameters optimization and feature selection and the improvement achievable by combining the neural network (NN) with an error imputation module (EIM) based on a regression trees ensemble	Shallow NN and Regression Trees Ensemble	30 and 60 min	RMSE, MAE, COD and delay	Y
Reddy et al [42].	2019	43	Energy expenditure (EE)	Even completely shutting insulin off at the time of exercise can still result in exercise-induced hypoglycemia.	Presented two new prediction algorithms with different levels of complexity to identify the risk of hypoglycemia at the start of exercise	Decision Tree, RF	NA	Accuracy, Sensitivity, Specificity, PPV, NPV, balanced accuracy, AUC	N
Romero-Ugalde et al. [43]	2019	15	The EE is computed from accelerometer and heart rate signals	It is very difficult to quantify the effect of PA on the physiological variables affecting the BG behavior.	Used variables usually modulated during and after a PA	ARX model	30, 60, and 120 min	IG prediction during 30 and 60 min (RMSE 7.75 $\pm$ 4.51 and RMSE 15.86 $\pm$ 9.61, respectively)	N

Sevil et al. [44]	2021	12	ML estimated the type and intensity of PA, the presence and characteristics of APS, and the concurrent presence of PA and Acute psychological stress (APS). Rule-Table for Fuzzy Logic Algorithm categorized as (SS: Sedentary State, DA: Daily Activities, RE: Resistance Exercise, TR: Treadmill Exercise, BK: Stationary Bike)	The mathematical relations between the physiological assessments and the glucose-insulin dynamics are complex and time-varying.	Integrated adaptive glucose prediction models with new features and metrics derived from biosignals to compute the effects of diverse PA and APS disturbances on GC predictions	Fuzzy Logic	1-hour, 1.5-hour, and 2-hour	MAE. Compared the results of the nominal model, which predicts the future glucose with only glucose and insulin information, to the nominal model + PA information, which incorporates an additional input representing the type and intensity of the PA	N
Shilo et al. [45]	2021	121	Daily activity (e.g., time from logged exercise)	Environmental factors, including the gut microbiota, are associated with the glycemic response of healthy individuals to meals	Constructing a prediction model for glycemic responses to meals specifically tailored for individuals with T1D using data on the administered insulin dosages, along with additional clinical and microbial data that were previously shown to contribute to Prediction of Personal Glycemic Responses in healthy individuals	XGBoost model	NA	SHapley Additive exPlanation (SHAP) methods	N
Sun et al. [46]	2021	NA	EE is estimated from physiological signals and is taken into account as one of the exogenous inputs while building prediction models.	Large amounts of data capturing the different states of a person are needed to train nonlinear models with a large number of model parameters.	Proposed a new PLS algorithm that incorporates regularization from prior knowledge and can handle missing data in the independent covariates	rPLS model with exogenous inputs	30 and 60 min	RMSE and MARD	N

Tyler et al. [47]	2022	20	Exercise history	Both automated hormone delivery and decision support systems currently lack the ability to accurately predict exercise-induced changes in glucose.	Personalized ML models were then designed to estimate the minimum glucose during aerobic exercise and 4 h following the start of exercise, and to quantify the impact of personalization on model accuracy. Considered three machine learning algorithms, a MARS, a logistic regression model, and an autoregressive (AR) model based on a previously described autoregressive model with exogenous inputs (ARX).	Multivariate Adaptive Regression Splines model to predict low Self-Monitoring of Blood Glucose (SMBG) after exercise, AR model to predict CGM following exercise, Logistic regression to predict hypoglycemia	NA	RMSE, MAE, as well as sensitivity, specificity, and accuracy to detect observations with level 1 hypoglycemia (< 70 mg/dL)	N
Vahedi et al. [48]	2018	93	Calories burned in heart rate, steps taken, EE	Having irrelevant and redundant features increases the computational complexity and decreases the performance of a predictive model	Sorted the features based on their correlation to Sensor Glucose. The highest accuracy is achieved by the best feature combination. The other method used the information gained to sort the features. Lastly, Principal Component Analysis was used to reduce the dimensionality of the dataset and find a set of new features with lower dimensionality that can provide better results.	Random Forest Regressor and MLP Regressor	NA	MAPE	N
van Doorn et al. [49]	2021	6	An accelerometer to assess physical activity	Large, human-based study populations are now needed to reliably assess to what extent and within what time interval (i.e., prediction horizon) glucose values can be accurately predicted by the use of machine learning.	A machine learning model that has been trained with a sliding time window of glucose values preceding the predicted values at a fixed interval. Additionally, whether glucose prediction can be further improved by the incorporation of accelerometer-measured physical activity was studied	Autoregressive Integrated Moving Average, Support Vector Regression, Gradient-boosting systems, shallow and deep multi-layer perceptron neural networks, and several recurrent neural network architectures, including classical RNN, gated recurrent units, long-short term memory networks, and all of its bi-directional variants	15 and 60 min	RMSE, Spearman's correlation coefficient (rho), and surveillance error grid	N

Vehí et al. [50]	2019	16	The activity on board is quantified based on the total steps performed by an individual throughout the day. The total number of steps performed over each sampling time is weighted by an exponential decay curve.	Improving the accuracy of BG level for patients with T1D	A system based on various methods of artificial intelligence for the prediction and prevention of hypoglycemic events in combination with data mining algorithms for the classification of glycemic control profiles for patients with T1D	ANN	30 and 60 min	K-fold cross-validation	N
Xie and Wang [51]	2020	6	5-minute aggregations of heart rate and the heart rate data were normalized to the scale of 0 to 1	There is no such a learning algorithm that outperforms any other learning algorithms in every problem set	Paper examines a set of machine-learning based regression models as well as two deep learning algorithms, by comparing their performance in predicting blood glucose levels	Elastic Net Regression, Gradient Boosting Trees, Huber Regression, Lasso Regression, Random Forest, Ridge Regression, Support Vector Regression (with Linear Kernel and Radial Basis Kernel, respectively) and Deep learning algorithm, LSTM and TCN	5 and 30 min	RMSE, TG	N
Zarkogianni et al. [52]	2014	6	Recorded EE of daily physical activities or exercise events from a wearable body monitoring system with a resolution time of 1 min.	The acceptance of CM models is limited because they take into account only a confined number of factors affecting the glucose metabolism and they are not easily individualized to accurately simulate metabolic processes for a specific Type 1 diabetes patient.	Applying neuro-fuzzy techniques taking input data from sensors for monitoring physiological parameters.	Neuro-fuzzy techniques while wavelets are applied as activation functions	15 min, 30 min, 45 min, and 60 min	RMSE and correlation coefficient (CC) corresponding to the testing datasets	N
Zarkogianni et al. [53]	2015	10	The sum of the energy expenditure during the time period [t—150 min, t—120 min] is fed into the models in order to take into account the physical activity during the latest 30 min with a lag time equal to 120 min.	Some of the endocrine processes affecting glucose metabolism are still not fully understood, these models take into account only a confined number of factors associated with glucose metabolism and cannot be easily individualized to accurately simulate metabolic processes for a specific T1DM patient	The use of data-driven modeling techniques has been proposed which disregard physiological insights and use pattern recognition techniques to simulate glucose metabolism	FNN, SOM, WFNN, and LRM	30, 60, and 120 min	RMSE, correlation coefficient (CC) and the mean absolute relative difference (MARD)	N

Zhang et al. [54]	2021	12	Work intensity, exercise duration, heart rate, steps, acceleration.	Data-driven personalized BG predictive models for T1D that are computationally efficient, suitable for wearable devices, and perform well on common problems that arise in the analysis of real-world data (missing data, uncalibrated data) for insulin therapy.	Applied efficient deep learning and regression models with/without encoder-decoder. Four data-driven models are presented, including two regression models and two deep neural network models.	Multiple Linear Regression model, Bidirectional reservoir computing model, Dilated convolutional neural network model, Sequence to sequence long short term memory model	30 and 60 min	RMSE, MAE	N

2.3.1 Summary of Review Table

- Study participants
  The study subjects were from all age groups, regardless of sex, the patients with T1D were counted as the study participants. Some studies used an on-request dataset such as OhioT1DM.

- PA input
  One of the most important strategies for successful hypoglycemia prediction is the selection of appropriate inputs. In most of the studies, CGM data or indices derived from CGM were used to predict hypoglycemia. In addition, other factors related to blood glucose levels, such as carbohydrate intake and physical activity, were also included. The studies with physical activity as an input are considered in this literature review, and the type of data treated as physical activity data is described here.

- Prediction Horizon (PH)
  PH is the amount of time the model has to predict the outcome in the future. There have been reports of prediction windows ranging from 15 minutes to four hours. It is natural to expect a decrease in predictive power as the PH increases. A shorter PH may be more useful for rapid clinician intervention, whereas a longer PH increases the prevalence of an outcome and consequently model performance, but may be less useful as a decision support tool due to the less accuracy. An increase in PH, on the other hand, may be useful for predicting BG after PA, as the effect of PA may take hours [55].

- Algorithm
  Various classes of ML techniques have been used in modeling such as neural networks (NNs), machine learning algorithms (e.g., Random Forest, Support Vector Machines), recurrent neural networks (LSTM), and various ensemble models.

- Validation approach
  Model validation is critical to developing algorithms and estimating performance. The most commonly used metric was RMSE, while others used AUC, F1 score, ROC curve, accuracy, sensitivity, MAE, MAPE, etc.

2.3.2 Summary of Modeling Approaches

To summarize the modeling approaches taken by each study, I have created a pie chart in Figure 2.2. The majority of the models fall under the category of ML models, which encompasses a broad range of algorithms and techniques such as classical regression models and an unsupervised model.

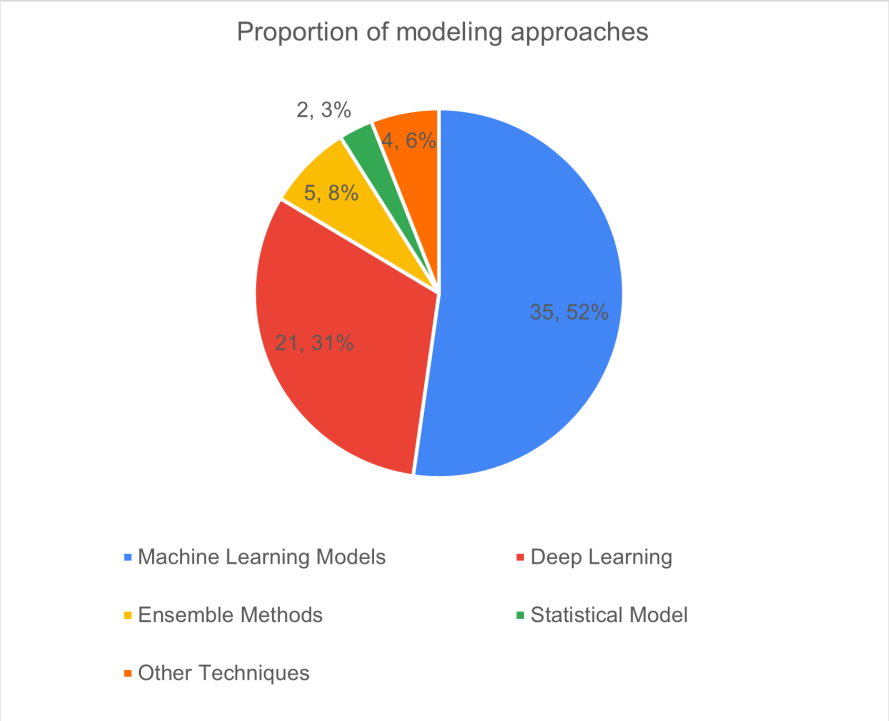

Figure 2.2: Models in Literature Review

The second most common category is Deep Learning, with 21 instances, featuring a variety of models including LSTM, ANN, and other variations of RNN.

2.4 Discussion

In this section, I explored the key factors that influenced the predictive models for BG levels, highlighting the significant insights from existing literature and the challenges that still remain.

2.4.1 Features used in training

Several studies emphasized the role of additional physiological and external factors in enhancing predictive accuracy. These included the incorporation of physiological signals (e.g., heart rate, oxygen saturation, motion), physiological impacts of insulin and carbohydrate absorption, and the effect of physical activities (PA) on blood glucose. Moreover, the impact of acute psychological stress (APS) on blood glucose dynamics was also explored in some studies [31] [32] [44].

2.4.2 Multitask Learning and Personalization

The effectiveness of multitask learning over single-task learning approaches was highlighted in certain scenarios, showing potential performance improvement across various prediction horizons [22]. Personalized models tailored for individual patients showcased consistency and performance improvements in terms of predictive accuracy [12] [19] [30] [41] [47] [53] [54].

2.4.3 Current challenges of using PA data in improving BG predictions

BG and PA are related in a very complex and interconnected way. The effect of PA in changing BG is related to the intensity and the duration of the PA. Providing accurate predictions of BG levels during and after PA is crucial for people with T1DM to effectively manage their health by avoiding hypoglycemia. However, incorporating PA into the prediction algorithm has not been an easy task due to the difficulty of collecting the data. They are requiring the patients to manually record data or resort to several sensors collecting data. These are prone to missing data for example, during the time charging the sensor, requiring an interpolation or other data preprocessing measures later on [56]. However, the interpolation approach has its own limitations. Imputation methods that involve both of the boundary points of the data gap (i.e., interpolation) cannot be used in a realistic online environment because one of the boundary points is in the future. Jeon et al. [32] explored different imputation methods on each patient's training set in the OhioT1DM cohort and compared test set prediction accuracy for each imputation method. They measured accuracy under two different conditions: a batch mode scenario (conventional train and test setting) and an online deployment setting (where future points are unknown).

The difficulty not only lies in the need to preprocess the PA data, rather a major difficulty is identifying what PA data is most relevant in the BG prediction algorithm. This is a crucial matter in building a prediction algorithm as irrelevant or redundant features can increase computational complexity and decrease model performance. [56] [57] claim that most available exercise models for T1D quantify exercise intensity or activity-related variables using percent oxygen consumption as a means to quantify exercise intensity.

On the other hand, Berford et al. [13] developed a Repeated Measures Linear Regression (RMLR) model, constructed using a generalized estimating equation (GEE) to estimate the model's parameters. Here the variable importance of the predictors for the final RMRF is found in the following order, glucose values at the start of exercise, glucose rate of change 15 minutes before the exercise, Insulin on board at the start of exercise, percent time <70mg/dL in the 24h before exercise, exercise start time, and etc.

The challenge still remains in the characteristics of PA and the behavior of BG. Tyler et

al. [47] demonstrate that even under highly controlled conditions, there is considerable intra-participant and inter-participant variability in glucose outcomes during and following exercise. Participants with higher aerobic fitness exhibited significantly lower minimum glucose and steeper glucose declines during exercise. Adaptive, personalized machine learning algorithms were designed to predict exercise-related glucose changes. In fact, the existing inter- and intra-patient variability in type 1 diabetes requires individualization of predictive models and their continuous adaptation to both biological and environmental changes [58] [59].

Another challenge of building a prediction model for BG level is the intrinsic nonlinearity and nonstationarity of the glucose regulatory system, nonlinear regression techniques of machine learning, such as feed-forward and recurrent neural networks, and Gaussian processes, have been used for predicting the glucose concentration in type 1 diabetes. However, nonlinear models come with a condition that many parameters require large datasets for training. Plus, deep learning models often lack interpretability. [19] and [45] tried to solve this issue by exploiting SHapley Additive exPlanations (SHAP).

2.5 Conclusion

Among the studies conducted addressing the identified challenges, researchers have defined various approaches, but commonly aiming to enhance the accuracy, personalization, and interpretability of blood glucose prediction models for individuals with Type 1 Diabetes. Based on the comprehensive analysis of these studies, future research directions could focus on further refining models through more sophisticated feature engineering, more comprehensive datasets, and additional physiological inputs. The exploration of more advanced models, possibly integrating both physiological and external factors in a comprehensive predictive framework, might offer more accurate and personalized predictions. Thus I plan to develop my own BG prediction model with the following specifications:

- Personalized predictive models
  To create personalized predictive models, machine learning models that take into account the individual variability in how T1DM patients respond to meals, insulin, physical activity, and other factors should be adapted by exploring and integrating multiple data sources, including continuous glucose monitoring (CGM), meal information, insulin dosing, and other vital signs.

- Physical Activity Integration
  Although the development of glucose level prediction algorithms has been studied for a long time, the number of prediction models that incorporate PA information is relatively small. Therefore, I decided to focus on the integration of physical activity

data into predictive models to address the challenge of accurately predicting glucose responses during exercise. The development of algorithms that can adapt to different types and intensities of physical activity should be explored.

- Missing Data Handling
  To improve the performance of the model, strategies are needed for handling missing data in T1DM prediction models, especially when dealing with data from continuous glucose monitors and other sensors.

- Multi-Modal Data Fusion
  Innovative methods need to be explored for integrating multiple types of data, including CGM, activity data, bio-signals, and insulin dosing information, to create holistic models for blood glucose prediction. The use of ensemble techniques or deep learning approaches for data fusion would be considered.

- Long-term Predictions
  If the scope of prediction models is to provide multi-step ahead predictions, the model needs to allow better long-term glycemic control. The use of recurrent neural networks (RNNs) or other time-series forecasting techniques should be investigated.

2.5.1 Direction of Research

With the information I have earned from this literature review, I have concluded to emphasize the uniqueness of my approach, by integrating physical activity data into blood glucose prediction models. By selecting and incorporating relevant physical activity data, my research aims to identify methods that enhance model performance and accuracy during physical activity, ultimately paving the way for more effective blood glucose prediction models for diabetes management.

3 Methods

In this chapter, I described the materials and software used throughout the study. The different methods employed for data collection is also explained, providing details about the data gathering process. Additionally, I outline the overall structure and planning of my study. Broadly, this thesis project can be divided into two parts: the comparison of blood glucose prediction models from five different papers with open-source code, and the study based on data collected from participants. Lastly the test plan for

3.1 Materials and Software

The study's computational framework is a critical component of the research process, providing the necessary resources to execute complex computations and deep learning tasks. In this section, the hardware and software environment that underpins the research is described.

3.1.1 Server Description

In this research, the computational work was performed using a dedicated server provided by the research group to which my thesis belongs. The server has 40 CPUs, and the L1d cache (640 KiB), L1i cache (640 KiB), L2 cache (20 MiB), and L3 cache (27.5 MiB) indicating high computational capacity and performance for repetitive tasks and complex computations.

This server served as the primary platform for training deep learning models.

3.1.2 Python Virtual Environments

Python virtual environments were used to isolate dependencies and create a consistent environment for different phases of the project, i.e. comparison work and working with real data using GluPredKit. The Python virtual environments ensured that the works were reproducible by avoiding conflicts between them.

3.1.3 Deep Learning in Python

The deep learning models in this study were designed and trained using TensorFlow, Keras, and PyTorch, frameworks that simplify building deep learning models and provide a variety of pre-built components.

Keras, a high-level API within TensorFlow, allowed the creation of sequential models with multiple types of layers, such as dense (fully connected) layers. Libraries for early stopping and learning rate adjustments were used to boost training efficiency and lower the risk of overfitting.

PyTorch, a deep learning framework with a flexible computation graph and extensive customization capabilities, was employed for more complex tasks. It played a key role in constructing Temporal Convolutional Networks (TCN) for this study. PyTorch also supports advanced training methods, including gradient clipping and learning rate scheduling.

Overall, the combination of TensorFlow, Keras, and PyTorch facilitated the development and training of deep learning models.

3.2 Data collection

This chapter outlines the processes involved in data acquisition and preparation for this research, focusing on the use of the OhioT1DM dataset and personal data collection from participants with type 1 diabetes. It also addresses the regulatory and ethical considerations involved in using personal data for research, as well as efforts made to contact authors of related papers to acquire additional resources.

3.2.1 OhioT1DM dataset

A request for the OhioT1DM dataset was made through Ohio University via my main supervisor, adhering to the strict requirements outlined by the dataset provider. The dataset is only available to established principal investigators affiliated with institutions engaged in research. Correctly completed forms were essential for the agreement process. It is noteworthy that the dataset is exclusively provided to employees of the institution, thus as a student researcher, I was required to request the dataset through my supervisor.

The form submitted to Ohio University comprises essential details, including the researcher's name, institutional mailing address, job title, and institutional email address. Moreover, an Institutional Contact, typically a legal signatory for the institution, has been specified. The Institutional Contact reviews and signs the agreement on behalf of the researcher. Ohio University then provides the agreement solely to the Institutional Contact for review and execution.

There was subsequent communication with Ohio University to clarify the progress of the request. Ohio University emphasized the submission of a fully executed Data Use Agreement (DUA) between the institutions, confirming compliance with the confidentiality and use terms outlined by Ohio University. This was signed by my main supervisor.

Apparently, the process involved multiple steps, including follow-up to ensure completion and receipt of required documentation. The dataset provider specified that the agreement was to be reviewed and executed only by the designated institutional contact, emphasizing adherence to protocol regarding data access and confidentiality.

Finally, the dataset was granted with specific instructions for accessing the OhioT1DM dataset, emphasizing encrypted storage and the need for a password to retrieve data. Additional resources were also provided, including a paper detailing the data format and a viewer for graphical data display.

3.2.2 Acquisition of Personal Data

The data collection process for this thesis project involves the acquisition of personal data from individuals diagnosed with type 1 diabetes. An initial inquiry was made to the UiT Personvernombud to seek advice and ensure compliance with data protection regulations. Below is a summary report detailing the communication initiated with the UiT Personvernombud regarding the use of personal data for this research.

Phase 1: Inquiry to UiT Personvernombud

The inquiry was sent to the UiT Personvernombud through an email to seek guidance on utilizing personal data for the development of a machine learning model focusing on individuals with type 1 diabetes during and after physical activity. The correspondence specifically sought advice regarding the potential implications of using personal data from supervisors who have agreed to provide the necessary information.

This interaction was initiated to align the data collection procedures with established regulations and ethical considerations, ensuring adherence to necessary guidelines in handling sensitive personal data for academic research purposes.

In conclusion, what was suggested by UiT Personvernombud is that I must register the project with SIKT.

Phase 2: SIKT Registration

All projects involving the processing of personal data must be registered with SIKT in accordance with UiT guidelines for research and student projects. Specific details about the project should be provided on the SIKT registration form. This includes information regarding the types of personal data to be processed, the data controller, the project details, samples, third-party access, documentation, approvals, security measures, closure, and additional project information.

As a part of the registration process, it is also required to fill out an information letter. The information letter serves as a comprehensive document outlining the project's purpose, responsible institutions, participant involvement, privacy measures, data storage, data ownership, and participants' rights.

The letter includes information on the purpose of the project, the responsible institution, reasons for participation, participant involvement, voluntary participation, data storage and usage, data retention, participant rights, the legal basis for data processing, and contact information for queries or rights exercises.

The information letter includes the consent form which will be provided to potential participants, giving them the opportunity to provide explicit consent for their participation in the project and the processing of their personal data. The consent form includes an agreement to participate, an understanding of the project, consent for data provision, and an agreement for the processing of personal data until a specified date.

While completing the information letter template, there were a few discussions with the thesis project supervisors. The email conversation involved discussions regarding the com-

pletion of an information letter template required for the registration of the Master's thesis project with SIKT. The communication mainly revolved around the template's details, including cooperation with specific institutions, data access, data retention, and supervisory roles.

I sought guidance from my supervisors regarding specific details to be included in the information letter. Answers were provided by supervisors and the things that were agreed upon were:

1. The institutes where co-supervisors are from should be mentioned in the template.

2. The co-supervisors from different institutes will have access to the data.

3. The data should be kept for 2 more years after the end of this project for verification and validation purposes of the study, also it may be used for future related projects.

The registration form was submitted to the website for review. It was mentioned that the review may take up to 1 month.

Phase 3: Assessment from Sikt

After a few modifications in the form, the assessment from Sikt regarding the processing of personal data in this project has been completed. The purpose of this assessment was to see if Sikt confirmed that the project has a legal basis to process personal data, aligning with data protection legislation. This assessment is in accordance with the agreement established between the institution where the student/researcher is affiliated and the Data Protection Services provided by Sikt.

Meanwhile, it also emphasizes adherence to the institution's guidelines for storing, transmitting, and securing the collected data. This necessitates utilizing data processors, such as cloud storage, online survey platforms, and video conferencing providers, with whom the institution has existing agreements.

The assessment conducted by Sikt is contingent upon meeting specific requirements outlined in data protection legislation, including accuracy (Article 5.1.d), integrity, confidentiality (Article 5.1.f), and ensuring overall security (Article 32) while processing personal data.

The assessment form states that any intentions to modify the processing of personal data within this project should be notified, by updating the information registered in the Notification Form. While ensuring that the changes align with the guidelines provided on the SIKT website, and awaiting Sikt's confirmation before implementing changes is advised.

Additionally, it plans to conduct a follow-up on the project at its planned end date. This follow-up aims to assess whether the processing of personal data has concluded as per the project's timeline and to ensure compliance with data protection regulations.

Phase 4: Defining data collection specifics

For the participant collected data, supervision meetings with my advisors focused on the type of data to collect. We reached a consensus that participants should not be confined to using a specific data collector. For instance, one participant utilized MiniMed for continuous glucose monitoring (CGM) and Fitbit for step counts, while another participant employed an Apple Watch, which captured CGM, insulin levels, carbohydrate intake, heart rate, and various other metrics. Furthermore, we agreed that the data collection does not need to adhere to uniformity in terms of the variables it captures. For instance, it is deemed acceptable as long as physical activity data is recorded in the format of steps, heart rate, calories burned, or a combination thereof, and the minimum required data for this study was included which are CGM, insulin, carbs, and physical activity data. However, unanimity was reached regarding the importance of collecting data at refined intervals, ideally between 5 to 15 minutes. The failure to do so could adversely affect the accuracy of predictions, even with preprocessing techniques such as imputation.

Steps, heart rate, and calories burned are frequently collected data points in health research due to their widespread availability through wearable devices and fitness trackers. They are considered indicators of physical activity (PA) in health literature and research studies because they provide direct measurements related to movement and energy expenditure. Steps count reflects the volume of ambulatory activity, heart rate indicates the intensity of physical exertion, and calories burned quantifies the energy expenditure associated with various activities [73]. By incorporating these metrics, individuals' engagement in physical activity, and monitor exercise intensity can be assessed.

While continuous monitoring devices are responsible for data acquisition, participants were advised to manually log their activities. This involves recording the start time, duration, and type of activity undertaken in the format of "YYYY-MM-DD HH:MM activity description." This method allows for the inclusion of activities not captured by the monitoring devices. This activity log would then be used during the evaluation of the trained model enabling the segmentation of the test data based on the activities documented in the log.

It was also decided that the data collection process would not adhere to strict controls, such as ensuring all participants engage in physical activity within 30 minutes of consuming carbohydrates. This decision stems from the recognition of the limitations in controlling such variables and the acknowledgment that data collected in free-living conditions may yield results closer to real-world scenarios. Although there were discussions about whether scenario-based data should be collected, such as a participant engaging in intensive activity

for 30 minutes every day, this idea was deferred for the same reason. The justification for collecting data in real-world settings is as follows:

Firstly, by gathering data in real-world settings, we replicate the natural environments and circumstances in which individuals with T1DM navigate their daily routines. This allows for a more reliable representation of the challenges and factors influencing blood glucose regulation during and after physical activities.

Secondly, real-world data collection enhances the generalizability of research findings by reflecting the diversity and variability inherent in real-life situations. Findings derived from studies conducted in controlled settings may not fully translate to the complexities of everyday life and the aim should be finding prediction models that can handle such underlying complexities.

Perception of Physical Activities

The following aspects will be considered as indicators of the physical activities of the participants:

- Heart rate levels:
  Heart rate can be used as a key physiological marker indicating the intensity of physical activity. As individuals engage in exercise, their heart rate typically rises to meet the increased demand for oxygen and energy. The degree of this increase correlates with the intensity of the activity [64].

- Step count:
  Step counts offer a quantifiable measure of movement and is widely recognized as a fundamental indicator of overall activity level. Incorporating step counts into the assessment could provide the volume and intensity of participants' daily physical activities.

Data Collection Period

The dataset will consist of three months' worth of data, including a week of physical activity logging. During the activity logging period, participants should record the details of their physical activities, noting the type and duration. At the same time, a smartwatch will automatically track heart rate and step counts. Participants are also encouraged to engage in a variety of activities that reflect typical daily routines, which will provide a more comprehensive view of real-world activity dynamics. The focus on capturing a wide range of activities enriches the dataset, making it more valuable for the study's overarching research objectives.

Sample of Activity Logs (Example)
The following exemplify a subset of the prescribed activity logging regimen mandated for all participants. This must be logged with the time that activities started and the time that the activities were completed:

- 2 sessions of snow shoveling, each lasting 20 minutes.
- 2 walking sessions, totaling 1 hour in duration.
- 1 ski trip, spanning a duration of 2 hours.
- 1 running session, with a duration of 30 minutes.

Rationale for Using Data Collected from Smart Watches

The use of smartwatches for monitoring physical activity and health parameters is justified, even though these devices are known to have some accuracy issues such as variations in heart rate monitoring and step counting.

Smartwatches offer a non-invasive and convenient way to continuously monitor various physiological metrics. This non-invasiveness improves user compliance and makes long-term data collection easier. Additionally, the participants with Type 1 Diabetes Mellitus (T1DM) were already accustomed to using their own smartwatches to track physical activity data.

The limitations inherent in smartwatches, including possible measurement inaccuracies and individual variations in data, are addressed in the Limitations section in the Conclusion chapter.

3.2.3 Attempt to Contact Authors of Relevant Papers

Attempts were made to acquire the source code or data necessary to validate the implementations derived from the referenced papers. I was particularly interested in TG, ESOD, and J index. I reached out via email to the authors of the related papers, believing that sharing their work and engaging in validation and potential improvements could offer value to the research endeavor. Unfortunately, I did not receive any response, which may be due to outdated contact information. Due to a lack of validation, I only explained the concept of TG, ESOD, and J index in the thesis.

3.3 Comparison Work

This comparative analysis primarily involves the replication of the reported experiments in the selected papers. Specifically, the validation approach includes obtaining the open-source code provided by each paper, setting up the necessary computational environment, and running the code to reproduce the reported results.

3.3.1 Goal of the Comparison Work

The purpose of this comparative analysis is to examine the existing works conducted by the authors of papers that utilize the OhioT1DM dataset. Considering the existing landscape found from the literature review, no single algorithm consistently outperforms others in terms of prediction accuracy. This analysis aims to validate some of the existing methodologies and gain insights throughout the exploration process. Also, it is to reveal the performance variations between different models, emphasizing their relative strengths and weaknesses. Additionally, this comparison aims to lay the foundation for synthesizing the most effective aspects of various algorithms, potentially leading to the development of hybrid or adaptive models.

3.3.2 Introduction of the Compared Papers

In this subsection, the papers selected for this comparison work are summarised. The detailed descriptions of each paper are introduced. Commonly the papers are published with open-source code, and all utilize the OhioT1DM dataset and share similar evaluation metrics, particularly the Root Mean Squared Error (RMSE). The following are the descriptions of each paper:

- Paper1 - **A Personalized and Interpretable Deep Learning Based Approach to Predict Blood Glucose Concentration in Type 1 Diabetes (2020)** by Cappon et al.
  - **Summary:** Utilized deep learning to predict blood glucose concentrations and exploited SHapley Additive exPlanations (SHAP) to interpret deep learning model predictions.
  - **Modeling approach:** Utilizes a bidirectional LSTM with multiple layers and a single neuron for BG prediction at different prediction horizons.
- Paper2 - **Data Fusion of Activity and CGM for Predicting Blood Glucose Levels (2020)** by Nemat et al.

  - **Summary:** Aimed to fuse continuous glucose monitoring (CGM) and activity data for blood glucose prediction. The research focuses on downsampling and fusion methods to generate models that are comparable to those using only CGM data.

  - **Modeling approach:** Three base regressions(MLP, LTSM, PSLR) and a stacked regression technique

- Paper3 - **LSTMs and Neural Attention Models for Blood Glucose Prediction: Comparative Experiments on Real and Synthetic Data (2019)** by Mirshekarian et al.

  - **Summary:** Investigated blood glucose prediction using recurrent neural networks (RNNs) based on simulated and real patient data. The comparison between different models and scenarios highlighted the accuracy of blood glucose level forecasting.

  - **Modeling approach:** Employed LSTM and a multitask learning approach in a neural network.

- Paper4 - **Personalised Glucose Prediction via Deep Multitask Networks (2020)** by Daniels et al.

  - **Summary:** Employed a multitask learning approach for glucose level prediction using neural networks. The study used a multitask learning (MTL) and single-task learning (STL) approach to improve glucose forecasting. Employed a multitask learning approach to improve the performance of glucose forecasting in a neural network, where each individual is viewed as a task, using shared layers to enable learning from other individuals.

  - **Modeling approach:** STL and MTL approach with CRNN Model.

- Paper5 - **Personalized Machine Learning Algorithm based on Shallow Network and Error Imputation Module for Improved Blood Glucose Prediction (2020)** by Pavan et al.

  - **Summary:** Focused on personalized machine learning algorithms to predict blood glucose levels. Investigated the impact of hyperparameter optimization and feature selection, particularly in combining a shallow neural network with an error imputation module.

  - **Modeling approach:** Utilized Shallow NN and Regression Trees Ensemble, focusing on feature selection and hyperparameter optimization.

Preprocessing Techniques for PA Data

The following is the description of how the preprocessing was handled in each paper.

Paper1

- Any missing data in the training set that was shorter than 30 minutes was filled using first-order interpolation. Signals from injected insulin, reported meals and physical activity are processed with a second-order low-pass filter (with a cutoff frequency of $\lambda = 0.02$) to capture the delayed impact on blood glucose (BG) levels. This approach was taken to account for the physiological dynamics, where effects are typically observed 30 to 60 minutes later.

Paper2

- In the training dataset, missing values were filled in using linear interpolation to maintain continuity and ensure data completeness. For the testing dataset, linear extrapolation was applied to simulate real-time scenarios, avoiding exposure to future data. This approach supports real-time applicability and maintains data consistency.

  To create a regular time series without missing values, continuous glucose monitoring (CGM) data was converted into 5-minute intervals, while activity data was transformed into 1-minute intervals. This consistent temporal structure provides a stable basis for analysis.

Paper3

- Any missing blood glucose level (BGL) values were filled in using linear interpolation. However, any data point where the target blood glucose value was derived from this interpolation was excluded from the dataset. Additionally, if there was a contiguous sequence of missing BGL values that ended at time, linear extrapolation was used to estimate these missing values.

Paper4

- It handled missing values by using linear interpolation for blood glucose data, imputed zeros for self-reported data, and scaled and transformed input features for training the deep multitask network. For example, it utilized a simple binary representation for exercise data, converting intensity values (1-10) into a presence or absence format, simplifying the information on exercise engagement.

Paper5

- In this paper, the preprocessing technique involves normalization and interpolation for handling missing values. Also, it applied a physical exercise-on-board feature using a second-order low-pass filter for intensity, which would likely capture a smoother representation of the exercise effect on BG levels.

Implementations of Each Paper

The following describes how the machine learning models are implemented in each of the five papers:

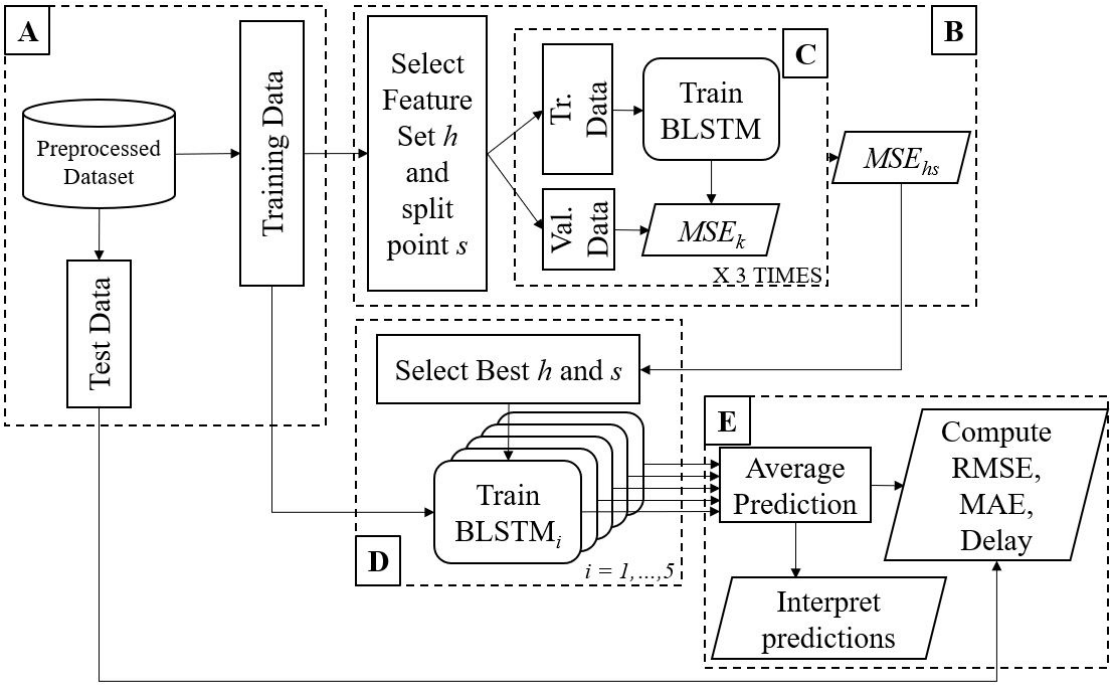

Figure 3.1: Framework of paper 1 [19]

In **Paper1**, the framework is structured into several blocks, detailing the process of training and evaluating BLSTM (Bidirectional Long Short-Term Memory) models for personalized blood glucose prediction. Block A describes the data preparation process. Data preprocessing involved dividing the data into training and test sets.

The Block B represents the feature selection and tuning process. Feature selection is performed by generating subsets of features, including the CGM feature, and assessing their impact on model performance. Data is split into training and validation sets using various split points (50%, 60%, 70%, etc.) to prevent overfitting.

Block C shows the model training process. The BLSTM's performance is evaluated for each feature set and split point in terms of mean squared error (MSE). To reduce the impact of random weight initialization, the training and evaluation process is repeated three times for each feature set and split point.

Block D is where the best feature set and split point are selected. The feature set and split point yielding the minimum MSE are chosen. Five BLSTMs are trained using the selected feature set and split point for specific patient/prediction horizons.

Block E represents the model evaluation phase. Model performance is assessed by comparing actual blood glucose (BG) values in the test set with predictions derived from averaging the estimates of the five BLSTMs. Model predictions are interpreted using SHAP (SHapley Additive exPlanations) to understand the model's decision-making process.

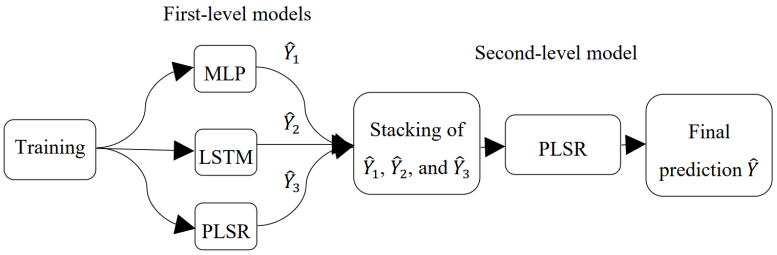

Figure 3.2: Framework of paper 2 [39]

The **Paper2** involved a stacked regression model aiming to improve blood glucose level predictions. Three base regression models (MLP, LTSM, PSLR) are utilized to generate initial predictions. Then, a Partial Least Squares Regression (PLSR) model is employed as a second-level model. This second-level model is trained using the predictions generated by the first-level models. To summarise, the predictions from the base regression models are used as features to train the second-level PLSR model.

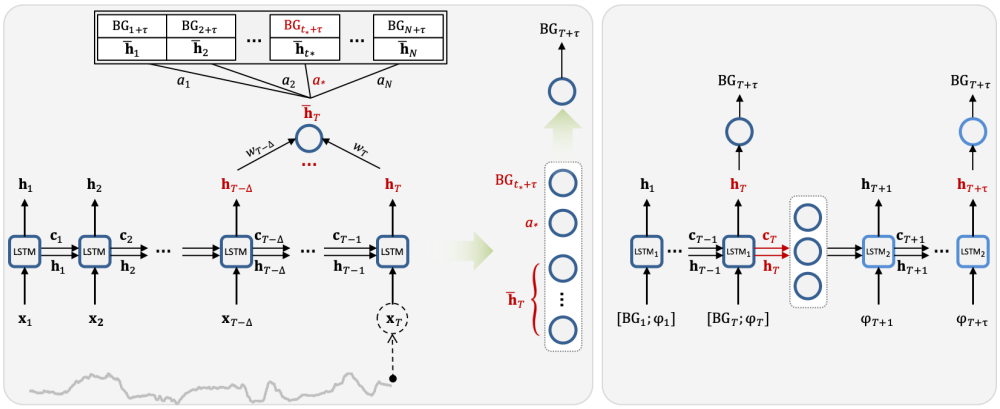

Figure 3.3: Framework of paper 3 [36]

In **Paper3**, Memory-Augmented LSTM (MemLSTM) comprises three key modules: LSTM, memory, and feed-forward. The LSTM module scans input, sequences of consecutive blood glucose level (BGL) readings, information on meals, insulin, and other activity data. Memory module consists of past h_t values and their corresponding target $BG_{t+\tau}$. The Feed-Forward module then computes the BGL prediction ($BG_{t+\tau}$) using information from LSTM and memory modules. Lastly, the attention mechanism utilizes maximum attention weight (a^*) aligning the LSTM state with memory content, aiding prediction.

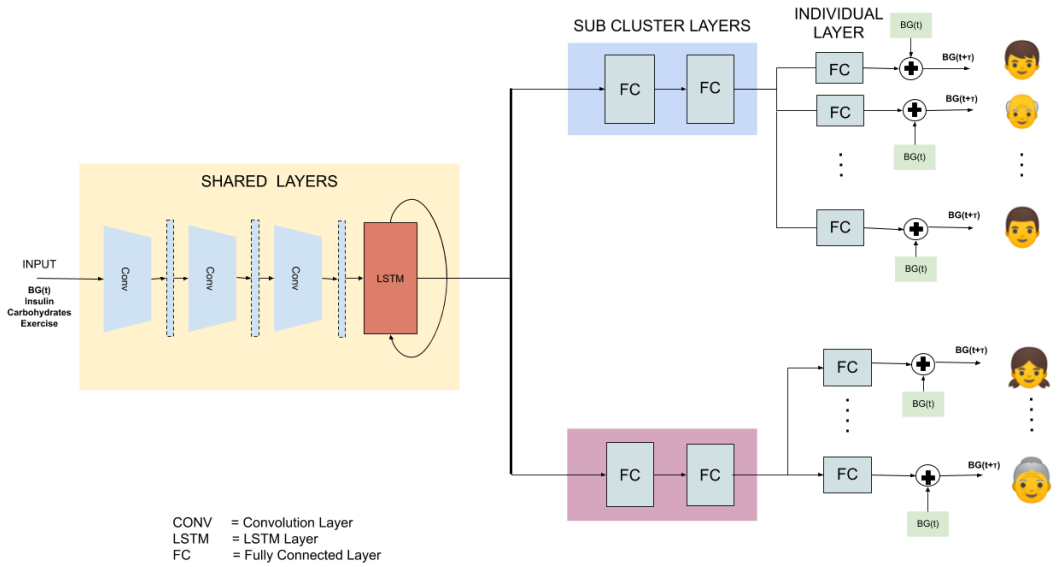

Figure 3.4: Framework of paper 4 [22]

The figure for **Paper4** illustrates the multitask learning approach used in personalized glucose prediction. Multitask learning aims to enhance generalization by simultaneously learning multiple tasks. In this context, each user is treated as a separate task. The initial shared layers produce outputs that are then fed into individual-specific fully connected layers. A multiplicative gating approach ensures that user-specific inputs are processed only within that user's individual-specific layers. During training, each batch contains data from a single individual, which trains both shared and individual-specific layers. The initial layers (including convolutional and recurrent) are shared among all users, the following two dense layers are shared based on gender, and the final dense layer is specific to each individual user.

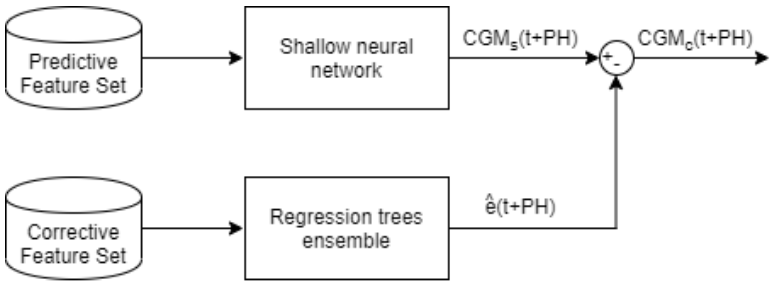

Figure 3.5: Framework of paper 5 [41]

The above figure shows how the model is designed in **Paper5**. The model comprises two primary parts, a Shallow Neural Network (NN) and an Error Imputation Module (EIM). The NN acts as the primary predictor, trained to forecast future blood glucose (BG) values for a specific prediction horizon (PH). The Error Imputation Module (EIM) functions as a predictor to estimate the error committed by the shallow NN.

The algorithm's prediction involves combining the shallow NN's output for CGMs(t + PH) with the EIM's output for $\hat{e}(t + PH)$. The resultant prediction incorporates both the original and corrected predictions for accurate glucose concentration estimation. In the Figure 3.5, $CGMs(t + PH)$ represents the original prediction, $CGMc(t + PH)$ denotes the corrected prediction, and $\hat{e}(t + PH)$ signifies the predicted error.

Challenges Identified by Each Paper

The selected papers present a variety of methodologies and approaches for predicting blood glucose levels in Type 1 Diabetes. These approaches employ different inputs and modeling strategies, aiming to enhance the performance of predictive algorithms. Each paper addresses distinct challenges within the field. The following challenges are presented in the order of **Paper1** to **Paper5**:

- Recurrent neural networks such as LSTMs are known to achieve good performance for the specific task of BG prediction, but they lack of interpretability.

- Data fusion of activity and CGM data normally results in models with a performance not comparable with those using CGM alone.

- When the blood glucose level is too high or too low, the individual reacts to bring it back into range but forecasting should enable people with T1DM to proactively intervene to prevent these conditions from occurring.

- Deep Learning models typically require relatively large amounts of data to converge on an appropriate model.

- Over the last two decades, several non-linear algorithms have been tested in this framework, but none has stood out in terms of prediction accuracy.

Proposal of Each Paper on How to Address the Challenge

Each paper proposed unique solutions to the challenges identified. They are presented in the sequence from Paper 1 to Paper 5:

- Utilized SHapley Additive exPlanations (SHAP), a newly developed approach, to interpret deep learning model predictions.

- Proposed two novel methods for fusing CGM and activity data to generate BGL prediction models with performance comparable to those using CGM data alone.

- Improved blood glucose level prediction using recurrent neural networks, leveraging both simulated patient data and data collected from individuals with T1D on insulin pump therapy with CGM.

- Employed a multitask learning approach to enhance glucose forecasting in a neural network, treating each individual as a task and using shared layers for learning from other individuals.

- Investigated the impact of hyperparameter optimization and feature selection, as well as the improvement achievable by combining the NN with an error imputation module (EIM) based on a regression trees ensemble.

3.3.3 Comparison Work Methods

Here is how the comparison process is carried out at each phase.

Step 1 - Validation with the Ohio T1DM Dataset

I started my analysis by examining the results from various research papers. These studies commonly used specific performance metrics—Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)—to measure predictive accuracy. Notably, they all reported results for 30-minute and 60-minute prediction windows. This consistency is because all the papers originated from the Knowledge Discovery in Healthcare Data 2020 event, which utilized the Ohio T1DM dataset. Using the same dataset provided a unified framework that facilitates comparative analysis among these studies.

Ensuring accurate outcomes required careful attention to several aspects. Achieving consistency across datasets, hyperparameters, and computing environments proved challenging. Some research papers lacked complete information, making it difficult to replicate their exact conditions. For example, while some code bases provided requirements.txt files that defined the Python environment, others did not include this detail. Resolving inconsistencies caused by conflicting modules or redundant dependencies in the requirements.txt files required considerable guesswork. Moreover, one paper added complexity by offering numerous adjustable hyperparameters without specifying the values used during their experiments, potentially leading to discrepancies between the actual and reported results.

Moreover, the utilization of MATLAB in one paper added another layer of complexity, requiring a licensed program. Luckily, access to MATLAB R2023b was facilitated by UiT, enabling the execution of the respective model on my local setup. Furthermore, harnessing the computational resources of our research group's server became imperative. Given the computationally intensive nature of the deep learning models employed, relying solely on my local system would have been suboptimal for executing these resource-heavy computations.

To ensure a comprehensive comparative analysis, a systematic approach was adopted to set up the environment and execute the provided open-source codes. The code implementations were carried out as per the instructions in the respective papers, with modifications kept to a minimum to maintain consistency across comparisons. Challenges encountered during the implementation phase were minimal due to the clarity and completeness of the provided code and documentation.

Step 2 - Validation with Participant Collected Data

In this section, the utilization of an external dataset extracted from Apple Watch is explored, to validate the predictive capabilities of the five prediction models from Knowledge Discovery in Healthcare Data and to allow the assessment of the generalizability and robustness of the prediction models across diverse health scenario. The dataset, comprised of 5-minute intervals, encompasses various health-related parameters recorded through Apple Health, providing a source of information for analysis. The Oura Ring data was also considered for testing but excluded due to data refinement challenges.

Step 3 - Implementing Prediction Models in GluPredKit

GluPredKit is an open-source comprehensive toolkit designed to facilitate the prediction of blood glucose levels [75]. It streamlines the process of data handling, training, and evaluating blood glucose prediction models in Python. These features are:

Data Parsing: It transforms data from different sources into a pandas DataFrame.

Data Preprocessing: It provides utilities for preprocessing the input data, which is crucial for building accurate prediction models. This step involves imputation, data scaling, feature selection, and train-test split to make data suitable for analysis.

Model Training: It has various types of prediction models for blood glucose prediction which provide implementations of various algorithms such as regression, classification, and time-series analysis to train predictive models.

Model Evaluation: The toolkit includes evaluation metrics, such as RMSE, to assess the performance of the trained models. This step is crucial for determining the accuracy and reliability of the predictions made by the models.

Overall, GluPredKit can serve as a resource for researchers and practitioners in the field of diabetes management, providing them with the necessary tools and methodologies to develop effective blood glucose prediction models.

Integrating GluPredKit into my master thesis will enhance the analysis and prediction aspects related to T1DM self-management during physical activities. The ease of conducting numerous tests and comparisons between different prediction models offered by the toolkit allows for comprehensive evaluation and validation of the predictive performance.

By incorporating GluPredKit into my analysis I expect to:

Improve Analysis and Prediction:
GluPredKit offers a framework for crafting customized prediction models specifically designed for predicting blood glucose levels. Utilizing its functionalities for data preprocessing, feature extraction, and model building ensures a consistent flow of preprocessing, model training, and evaluation. This streamlined approach applies consistently across various types of data supported by the framework, thereby enhancing the accuracy and reliability of blood glucose predictions.

Facilitate Comparative Studies:
The ability to conduct comparisons between different prediction models within GluPredKit allows for a thorough evaluation of their performance. This comparative analysis can help identify the most effective model or combination of models for predicting blood glucose levels in varying scenarios.

Overall, integrating GluPredKit into my master thesis not only strengthens the analysis and prediction aspects related to T1DM self-management but also provides a platform for conducting rigorous comparative studies.

3.4 Working with Participant Collected Data

This section describes the machine learning models, evaluation metrics, and evaluation intervals used during the phase of working with participant-collected data.

3.4.1 Evaluated Models

This subsection examines a variety of machine learning models, including linear regression models, tree-based models, neural networks, and support vector machines, each with its unique characteristics and applications. These models serve as the foundation for other approaches, such as ensemble models and physiological hybrid models, which will be discussed in detail in the Method chapter.

Linear regression models

Linear regression is based on the assumption that the relationship between the variables is linear, allowing for straightforward analysis and prediction. However, it has limitations when the data exhibits complex or non-linear patterns.

ARX (Auto-Regressive with Exogenous Input) ARX models are used to predict a time-series target variable by considering its past values (auto-regression) and additional

exogenous inputs. Exogenous inputs refer to other independent variables that could affect the target variable, such as insulin, carbohydrates, and other physical activity-related variables. in GluPredKit, ARX is implemented using Scikit-learn's LinearRegression which fits a linear model with coefficients to minimize the residual sum of squares between the observed targets in the dataset and those predicted by the linear model [61].

Elastic Net Elastic Net is a type of linear regression that brings together the key elements of both Lasso and Ridge regression. It includes a penalty term that blends the L1 penalty from Lasso with the L2 penalty from Ridge. Because it uses both penalties, Elastic Net can capture the advantages of each approach.

Huber Huber Regression is a type of L2-regularized linear regression that can handle outliers better than traditional linear regression. It uses two types of loss functions: squared loss for smaller errors (to maintain precision) and absolute loss for larger errors (to be robust to outliers). It includes parameters for optimizing the model's weights, intercept, and scale. The scale parameter (sigma) ensures consistent robustness even when the target variable changes its range or scale. The advantage of the Huber loss function is that it doesn't get overly skewed by outliers while still accounting for their impact [62].

Lasso Lasso, which stands for "Least Absolute Shrinkage and Selection Operator," is a kind of linear regression that applies L1 regularization to impose a penalty on large coefficients. The aim is to enhance model accuracy by reducing overfitting and by choosing the most significant features.

PLSR (Partial Least Squares Regression) Partial Least Squares Regression (PLSR) is a method that integrates aspects of principal component analysis and multiple linear regression. It is used to understand complex relationships between a set of independent variables and a dependent variable, particularly when the predictors are numerous and might be strongly correlated. This is achieved by projecting the independent variables and dependent variables into a new, lower-dimensional space. The approach creates new components that capture the most variation in the variables, helping to manage large sets of correlated variables while minimizing multicollinearity issues.

Ridge Regression Ridge regression is a type of linear regression that includes a regularization term to prevent overfitting. It achieves this by adding a penalty to the loss function, which is calculated as the sum of squared coefficients (L2-norm). The regularization term controls the trade-off between fitting the training data and keeping the model coefficients small, reducing the risk of overfitting to noise or fluctuations in the data.

Tree Based Models

Tree-based models use a tree-like structure for decision-making and prediction. At each node, the model makes a decision based on specific conditions or features, leading to branches that represent different outcomes. The process continues until reaching the leaves, which represent the final predictions.

GBT (Gradient Boosting Trees) Gradient Boosting Trees (GBT) begins with a basic decision tree, a model that makes choices based on defined rules. Each node in the tree represents a decision point, while the leaves signify the final outcomes. The boosting process involves adding additional models to correct the mistakes of earlier ones, building trees one by one, each concentrating on the errors made by the previous trees. This method is designed to reduce errors by tweaking the models based on how much they deviate from the correct predictions. It essentially finds the optimal way to minimize mistakes. The final prediction comes from a group of these smaller "weak" models, typically decision trees. For regression tasks, the end result is often calculated by averaging the outputs of all the trees, leading to improved model performance.

Neural Network Models

Neural Network Models consist of interconnected nodes, or neurons, arranged in layers, with each layer performing a specific role in the learning process. Neural networks can learn non-linear relationships, making them highly adaptable and suitable for complex tasks.

LSTM (Long Short-Term Memory) Long Short-Term Memory (LSTM) networks are a special type designed to excel at handling sequential data. This can also include time series information on blood glucose levels, where the order and context of the data points are crucial. LSTMs achieve this by overcoming the vanishing gradient problem, a limitation that hinders standard RNNs from capturing long-term dependencies within the data.

TCN (Temporal Convolutional Networks) Temporal Convolutional Networks (TCNs) are a kind of neural network designed to work with data that has a sequence, like how Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks do. But instead of using recurrent connections like RNNs, TCNs rely on convolutional layers to process sequences. This means they apply filters over the sequence to extract information. The core structure of a TCN is one-dimensional, and it uses padding with zeros to ensure the output sequence remains the same length as the input sequence [63].

MLP (Multi-Layer Perceptron) A Multi-Layer Perceptron (MLP) is a foundational type of artificial neural network. These MLPs are characterized by their layered structure,

typically consisting of an input layer, one or more hidden layers responsible for the "multi-layer" aspect of the name, and finally an output layer. Each layer within this network houses multiple interconnected processing units known as neurons.

Support Vector Machines

Support Vector Machines (SVMs) are a type of supervised learning algorithm used for both classification and regression tasks. The core concept behind SVMs is to find the optimal hyperplane, a decision boundary, that separates data into distinct categories, with the goal of maximizing the margin between different classes.

SVR (Support Vector Regression) Support Vector Regression (SVR) is a type of machine learning used to forecast continuous values. Instead of fitting a straight line through the data like traditional linear regression, SVR identifies a hyperplane in a multi-dimensional space that best represents the data. This method helps handle complex relationships and reduces overfitting by concentrating on the support vectors, which are the data points nearest to the hyperplane. The objective is to keep most data within a defined margin around the hyperplane while minimizing errors.

3.4.2 Evaluation Metrics

This subsection explores key metrics and visualization tools used to evaluate the performance and reliability of blood glucose prediction models. It includes several methods for measuring error, visualizing prediction accuracy, and assessing clinical risk.

RMSE (Root Mean Squared Error)

RMSE is calculated by taking the square root of the average of the squared differences between the predicted values and the actual values. It penalizes large errors more heavily than small errors due to the squaring operation. It provides a measure of how spread out the errors are in the predicted values.

RMSE during PA

The RMSE during the physical activity (PA) period was calculated to evaluate the accuracy of predictions during these times. This metric was calculated for each activity's duration, from the start time to the end of the activity. The detail of the code is as shown in Listing 3.1.

Listing 3.1: Part of the RMSE during PA calculation function

```
rmse = 0
rmse_list = []

# Iterate through activity logs
for log in activity_logs:
    start_time = pd.to_datetime(log['start_time'])
    duration = log['duration']
    end_time = start_time + pd.Timedelta(minutes=(duration)
       * 5)

    # Find indices within the activity period
    indices = (y_true.index >= start_time) & (y_true.index
       <= end_time)

    # Calculate RMSE for the activity period
    rmse = np.sqrt(np.mean(np.square(np.array(y_true)[
       indices] - np.array(y_pred)[indices])))
    rmse_list.append(rmse)
```

MAE (Mean Absolute Error)

MAE is calculated by taking the average of the absolute differences between the predicted values and the actual values. It treats all errors equally, regardless of their magnitude, making it less sensitive to outliers compared to RMSE. It provides a measure of the average magnitude of errors in the predicted values.

Since the blood glucose level data contains outliers that could have a significant impact on the model's performance evaluation, RMSE is appropriate so that it penalizes large errors more heavily. It is indeed widely used to assess the performance of the BG prediction models. However, I have also evaluated the models with MAE to reflect the average magnitude of errors without any transformations. During the evaluation, I observed that the MAE score is usually lower than the RMSE. This is an expected observation considering that RMSE penalizes larger errors more heavily due to the squaring operation.

nRMSE (Normalized Root Mean Squared Error)

Normalized RMSE is used to standardize the scale of RMSE scores, allowing for comparison across participants and activity periods. In this study, one participant frequently exhibited extreme hyperglycemic values, while the other generally maintained blood glucose within the safe range of 70-180 mg/dl. Additionally, within a single individual, blood glucose levels can fluctuate throughout the day, particularly during different physical activity periods. By

normalizing the RMSE, the scale of the evaluation metric is standardized, facilitating easier comparison of model performance across these varied conditions.

To normalize the RMSE, I considered four approaches: using the interquartile range (IQR), standard deviation, mean value, and the range (difference between maximum and minimum). Ultimately, I chose to use standard deviation for normalization to address variations in RMSE values during physical activity. The standard deviation was calculated over the observed data. However, I chose not to use normalized RMSE in the final analysis because raw RMSE appeared to better capture fluctuations in the model's predictions within a single participant's data. Despite this, I believe that normalized RMSE could be useful in future studies, especially when comparing results across a larger number of participants to assess variability among them.

Trajectorty plot

A trajectory plot is a visualization tool used in machine learning and data analysis to assess the performance and behavior of a model over time or across different conditions. It typically involves plotting the predicted values against the actual values or some other relevant variable.

Trajectory plots are valuable for analyzing model performance as they provide a visual representation of how well a model's predictions align with the actual values or expected trends in time series or longitudinal data. This visual insight can often reveal patterns or discrepancies that may not be immediately apparent from numerical metrics alone, such as RMSE or MAE. Trajectory plots can illustrate how a model's predictions evolve over time. This can be crucial for detecting trends or shifts in the underlying data patterns.

Trajectory plots can be used as a diagnostic tool to identify areas where a model may be performing well or struggling. Deviations from the ideal trajectory can indicate areas for improvement or further investigation.

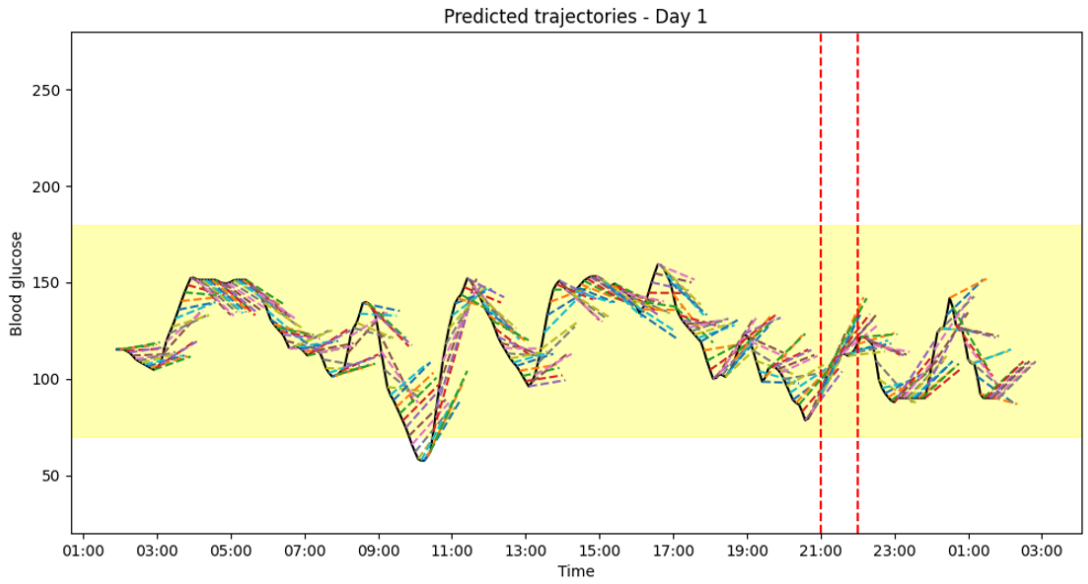

Figure 3.6: A ridge model trajectory plot with the 60-minute prediction horizon for participant 1

The figure above illustrates the trajectory plot of the Ridge model trained using Participant 1's data. The trajectory plots are drawn for the periods that the participants logged their PA. The plot is drawn per day. The solid black line represents the actual blood glucose (BG) levels, while the shaded yellow background signifies the safe BG range between 70 to 180 mg/dL. Deviations below or above this range indicate hypoglycemia or hyperglycemia, respectively. A red dotted line denotes the period of physical activity, emphasizing the model's performance during this phase. Colored lines extending from the black line depict the Ridge model's predictions for the subsequent 60 minutes, offering insights into anticipated BG fluctuations. A closer alignment between these projected lines and the black line signifies accurate predictions, whereas skewed trajectories indicate potential prediction discrepancies. This visualization aids in assessing the model's predictive efficacy, crucial for optimizing self-management strategies in individuals with T1DM.

By incorporating trajectory plots into this thesis work, I expect to interpret how well the model captures the underlying relationships in the data. This interpretability is especially valuable in fields where understanding the model's behavior is as important as predictive accuracy. Also, I aim to illustrate the performance of each model evaluated provides valuable insights into their behavior and the effectiveness of the prediction outcome.

Clarke Error Grid

The Clarke Error Grid Analysis (EGA) was introduced in 1987 as a method for assessing the clinical accuracy of patient estimates of their current blood glucose relative to the values obtained from their meters. [60] Subsequently, it has been applied to evaluate the clinical accuracy of blood glucose estimates produced by meters in comparison to a reference value. It is a graphical tool to evaluate the accuracy of blood glucose measurement. It plots the relationship between measured blood glucose values and clinically relevant outcomes, such as the risk of hypo- or hyperglycemia.

In the CEG, the x-axis represents the blood glucose values, while the y-axis represents the CGM blood glucose values. Then the grid is divided into zones that indicate different levels of clinical significance.

Zone A represents clinically accurate measurements where values fall within a predefined acceptable range. Measurements in this zone are considered clinically safe. Zone B represents measurements that are clinically acceptable predictions but may lead to benign errors in treatment decisions. Measurement in this zone can be interpreted that it is slightly outside the acceptable range but does not result in significant clinical consequences. Zone C, D, and E represent increasingly significant discrepancies between blood glucose values. Zone E indicates the most critical errors that could potentially lead to a dangerous treatment decision.

It provides a visual representation of measurement accuracy and thus can assess the reliability and safety of the method being evaluated.

Parkes Error Grid

Parkes Error Grid Analysis is another graphical tool used to evaluate the accuracy of blood glucose prediction. The Parkes error grid was published in 2000 based on a survey of 100 physician attendees at the June 1994 American Diabetes Meeting. Similar to the Clarke Error Grid, the Parkes Error Grid plots the relationship between measured blood glucose values and clinically relevant outcomes and specifies five risk levels. The interpretation of each zone is also similar.

Parkes error grids were made for two groups: people with type 1 diabetes and people with type 2 diabetes who use insulin. The idea came from feedback from Parkes survey participants, who thought that people with type 2 diabetes using insulin could handle more mistakes in blood glucose readings than those with type 1 diabetes. Thus the type 2 diabetes grid was seen as needing less accuracy, especially for low glucose levels, compared to the type 1 diabetes grid.

3.4.3 Evaluation Interval

In this subsection, I explained the prediction horizon selected for evaluating the models and the reasons behind this choice.

Prediction Horizon

The selection of a prediction horizon for blood glucose (BG) prediction models commonly falls within the 30-60 minute range, primarily due to the inherent volatility of BG levels. Various factors, including food intake, physical activity, stress, and insulin administration, can trigger rapid fluctuations in BG [69]. However, the rate of these fluctuations tends to decrease over time [70]. Consequently, the 30-60 minute window captures significant BG changes, making it a suitable timeframe for prediction.

Extending the prediction horizon beyond 60 minutes often necessitates more intricate models that incorporate longer-term trends and dynamics of glucose metabolism. In the context of forecasting blood glucose levels in individuals with type 1 diabetes, longer prediction horizons may involve predicting how glucose levels will evolve over a longer period, which can be influenced by various factors such as meal intake, insulin dosing, physical activity, and other physiological variables [71]. Models like the CNN-LSTM stacked architecture mentioned in the paper are designed to capture these intricate dynamics and trends to make accurate predictions over longer time horizons. Such models can offer valuable insights into future BG patterns but frequently entail increased computational complexity.

Furthermore, individual responses to various stimuli can exhibit substantial variability, potentially rendering longer-term predictions less reliable for certain individuals.

3.4.4 Data Considerations

For participant data collection, several sources were considered, and ultimately Apple Health, Medtronic MiniMed, and Fitbit were chosen. Below is an overview of the data formats from these sources.

Oura Ring Data Consideration

The Oura Ring, renowned for its advanced sleep tracking, activity monitoring, and physiological metrics, presented an opportunity to enrich the dataset further. However, during the exploration phase, it was observed that the Oura Ring did not provide the functionality to export data at 5-minute intervals. Unfortunately, this limitation rendered the data less

refined for the prediction model's requirements. The model's success relies on detailed temporal patterns, and without this level of granularity, the Oura Ring data did not align with the specificity required for testing the models.

Apple Health Data Overview

The apple_health.CSV file extracted from Apple Health contains the following key columns:

- date: Timestamps indicating the date and time of each record.
- CGM (Continuous Glucose Monitoring): Data related to glucose levels, measured at 5-minute intervals.
- carbs: The amount of carbohydrates consumed at each 5-minute interval.
- insulin: Information about insulin intake at each 5-minute interval.
- heartrate: Heart rate measurements recorded at 5-minute intervals.
- heartratevariability: Data related to heart rate variability, a measure of the variation in time between heartbeats.
- caloriesburned: Information about calories burned during each 5-minute interval.
- respiratoryrate: Data on respiratory rates, indicating breaths per minute.
- vo2max: Representation of the maximum rate of oxygen consumption during exercise.
- steps: The number of steps taken in each 5-minute interval.
- restingheartrate: The resting heart rate, providing a baseline for cardiovascular health.
- activity_state: Indication of the state of physical activity during each 5-minute interval.
- hour: The hour component of the timestamp.

Medtronic MiniMed Data Overview

Data collected includes information sourced from the Medtronic MiniMed™ 780G insulin pump, which provides non-periodic data intervals for bolus and carbohydrate intake, recorded manually. Blood glucose levels are recorded every 5 minutes. Consequently, the

exported data consists of distinct sections, with one containing non-periodic records and another dedicated to blood glucose levels. This dataset is then integrated with step count records from Fitbit devices.

Fitbit Data Overview

The process by which Fitbit converts accelerometer data into steps is not openly disclosed. It can be inferred that like most the activity monitors, it seeks specific motion patterns that meet a detection threshold indicative of walking [74]. If the pattern and magnitude of the motions align with the algorithm's predefined criteria, they are counted as steps.

The extracted step count data was logged every minute, but it needed to be processed to align with the format of the other features, which were logged every 5 minutes. Therefore, only the data from each 5-minute interval was collected and used for training.

3.4.5 Data Parsing

Participant 1's data were given in two files. One of these files was an exported CSV file from MiniMed 780 MMT-1885. The data within this file was divided into various sections, and the specific data required for analysis was scattered across these sections. Parsing this data presented a challenge due to the fluctuating format with each extraction. It is suspected that this variability was intentional, possibly designed to safeguard the data from integration with other systems. Consequently, the parsing process had to be adapted for each data extraction.

The second file contained data in JSON format, comprising step count measurements from Fitbit taken at five-minute intervals. These measurements were aggregated into a single CSV file, which served as the basis for preprocessing.

Participant 2's data were provided in parsed CSV format, facilitating straightforward preprocessing using the built-in methods of GluPredKit.

3.4.6 Data Preprocessing

There are different types of preprocessing readily available in GluPredKit but they are variations of each other but basic preprocessing principles are similar.

Initially, the preprocessor selects the relevant numerical and categorical features from the DataFrame came from parsing the raw data. It adds a target column by shifting the CGM (Continuous Glucose Monitoring) values backward by a certain number of steps, determined

by the prediction horizon.

Before proceeding with further processing of the data, any NaN values in the CGM column were checked to mark them 'imputed' in a separate column to mark instances where imputation is needed. For the imputation tactic, it adopted the Akima interpolation technique to handle missing numerical values. The Akima interpolation uses local methods that use only values from neighboring knot points in the construction of the coefficients of the interpolation polynomial between any two knot points thus it can be calculated very quickly [76].

If numerical features are present, the values are scaled using the StandardScaler, fitting the scaler on the training data and then transforming both training and testing data. For categorical features, they are one-hot encoded using the OneHotEncoder.

Overall, the preprocessing pipeline ensured that the data was properly formatted, imputed, scaled, and encoded, making it suitable for training machine learning models.

3.5 Approaches Taken to Improve the Predictions During PA

In this section, I have explained how the two distinct approaches to building models for blood glucose prediction are implemented: hybrid physiological-machine learning (ML) methods and ensemble models. The intention of these approaches was to see if they improve the predictions during the physical activities (PA). Additionally, the motivations for adopting these methods are discussed.

3.5.1 Hybrid Model Combining Machine Learning and Physiological Principles

A physiological model encompasses a mathematical description of the intricate biological processes within the human body, focusing on aspects related to health and well-being. Specifically within the realm of diabetes management, these models are crafted to simulate the complex interplay between variables such as plasma glucose, insulin, and carbohydrate levels, which are influenced by processes including digestion, absorption, insulin-dependent and independent utilization, renal clearance, and endogenous liver production [83]. Rooted in the comprehension of human physiology, these models aim to forecast how these variables interact to shape blood glucose levels over time, offering a robust framework grounded in scientific principles and established physiological mechanisms. While providing valuable insights into the impact of factors such as insulin, carbohydrates, and physical activity on

blood glucose levels, these models may fall short of capturing the full spectrum of individual responses to diverse stimuli.

Machine learning (ML) models, in principle, should be capable of identifying these patterns without transformation even in the absence of explicit understanding or characterization, provided there is a sufficient volume of data and accurately recorded input signals. However, it is common practice to engage in "feature engineering" to translate model inputs into physiological dynamics [87]. For instance, in the case of insulin, this may involve calculating "insulin on board" as opposed to utilizing time-lagged features. Such transformations may aid ML models in more readily discerning patterns, as the input becomes more directly correlated with the output.

Another approach, blending elements of both physiological modeling and ML, involves initially employing a model-based predictor, such as physiological models for insulin and carbohydrates (and possibly physical activity) [41]. Then subsequently, ML models can be utilized to predict the error of the physiological model rather than directly forecasting blood glucose levels. This approach capitalizes on the understanding humans have regarding insulin-glucose dynamics, while allowing ML to capture more intricate patterns, thus presenting a synergistic fusion of knowledge-driven physiological insights and data-driven ML capabilities.

On the other hand, an innovative approach inspired by metabolic models for glucose dynamics [83] [85] [86] is introduced by Mario Munoz-Organero [84]. This novel mechanism is designed to be trainable on a per-patient basis, aiming to capture the complex interactions between various factors influencing blood glucose levels. Specifically, the model leverages Recurrent Neural Network (RNN) architecture, implemented using Long Short-Term Memory (LSTM) cells, to simulate the differential equations governing carbohydrate and insulin absorption processes.

The dynamics of blood glucose levels over time are influenced by multiple factors, including current blood glucose levels, carbohydrate intake, and insulin injections, each characterized by specific absorption rates. These processes are modeled using a set of differential equations previously proposed in research studies [83] [85] [86], accounting for digestion, absorption, insulin-dependent and independent utilization, renal clearance, and endogenous liver production processes.

The paper explains that by considering inputs such as carbohydrate intake, fast and slow-acting insulin boluses, and past blood glucose levels, the RNN is trained to learn the digestion and absorption processes. The model learns from current values and past data, capturing the temporal dynamics of these processes. The proposed method involves a two-layered architecture. Initially, the LSTM RNN learns the carbohydrate digestion and insulin absorption processes from each input signal. Subsequently, the effects of these processes are combined to predict blood glucose variations for the next Continuous Glucose Monitor

(CGM) reading.

The universal approximation theorem [88] suggests that with enough hidden neurons, an RNN can approximate a physiological model, regardless of the state transition equations it uses.

Physiological Hybrid Approach Implementations

The deep physiological model [84] is implemented using a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells. This implementation is motivated by the need to mimic the metabolic behavior of physiological blood glucose models while leveraging the power of deep learning techniques.

Traditional physiological-metabolic models have limitations in capturing the intricate relationships between various factors influencing blood glucose levels, such as carbohydrate intake, insulin injections, and individual metabolic responses. By integrating deep learning components like RNN with LSTM cells, the model is expected to learn and adapt to the unique characteristics of each patient, providing more personalized and accurate predictions.

The model in Figure 3.7 uses separate Long Short-Term Memory (LSTM) networks to understand different types of data related to blood glucose (BG) levels. One LSTM network is designed to analyze the patterns in the blood glucose readings over time. Another set of LSTMs is used to process data related to insulin intake and carbohydrate consumption. The outputs from these different LSTMs are combined to create a more comprehensive representation of the data. By integrating these outputs, the model aims to predict the expected change in blood glucose levels for the next continuous glucose monitoring (CGM) reading. In simpler terms, the model uses multiple LSTMs to process different types of inputs, and then combines their outputs to make a prediction about future blood glucose levels.

3.5.2 Ensemble Models

Many papers investigated ensemble methods in multiple medical subfields for both classification and regression tasks [77] [78] [79] [80] [81]. The application of ensemble models can provide several advantages, including enhanced accuracy through the aggregation of diverse base learners, each adept at capturing distinct aspects of the data. Furthermore, the characteristic of robustness to noise and outliers can render ensemble methods particularly suitable for the blood glucose level prediction model. Through the principle of consensus learning, wherein multiple models collectively contribute to the final prediction, ensemble methods circumvent individual model biases, further improving prediction quality. To sum-

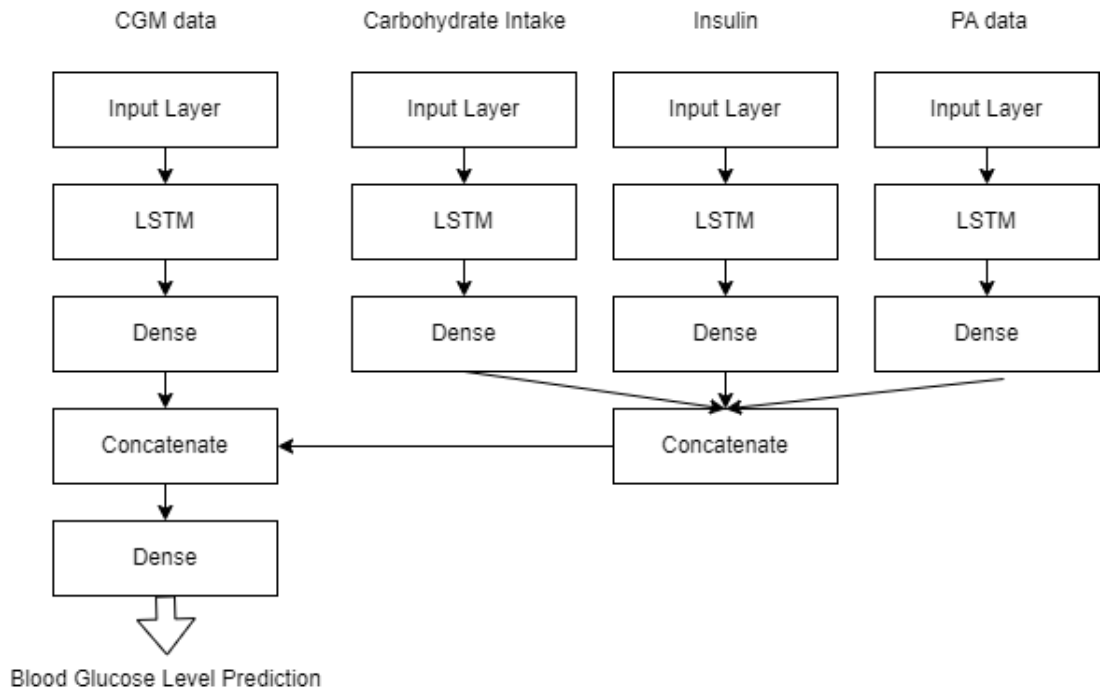

Figure 3.7: Model inspired by the Deep Physiological Model [84]

marise, by leveraging the collective intelligence of multiple models, it is expected that the ensemble methods will offer a powerful approach to predicting blood glucose levels with better accuracy and reliability.

Types of Ensemble Models

Ensemble learning broadly falls into two categories: homogeneous and heterogeneous ensembles.

Homogeneous Ensembles:
Homogeneous ensembles refer to methods in which all base learners share a similar structure or design. In bagging, multiple base learners are trained on different subsets of the training data, created by sampling with replacement. The final prediction is usually derived by averaging the predictions of all base learners. A well-known example of a bagging-based ensemble is Random Forest.

Boosting is another notable homogeneous ensemble technique, distinguished by its iterative approach. It involves training base learners sequentially, with each learner concentrating on correcting the errors made by its predecessor. This concept is embodied in algorithms

like AdaBoost, Gradient Boosting, XGBoost, and LightGBM.

Stacking, also known as Stacked Generalization, is a method where different base learners are trained, and their predictions are used as input features for a meta-learner. The meta-learner then combines these predictions to generate the final output.

Heterogeneous Ensembles:
Heterogeneous ensembles differ by combining base learners of various types, typically from different machine learning algorithms. By leveraging the unique strengths of these diverse learners, heterogeneous ensembles aim to improve prediction performance. For example, this approach might pair decision trees with neural networks or support vector machines in the same ensemble. The goal is to encourage diversity among the base learners, enhancing the ensemble's overall predictive capability.

Homogeneous and heterogeneous ensembles both aim to reduce overfitting, improve generalization, and boost prediction accuracy by combining the outputs of multiple models. The choice between these two types of ensembles depends on various factors, including the nature of the data, the specific problem being addressed, and the overall goals of the prediction task. However, there is no standard process to select the base learners and it still remains as a challenge for researchers to find the best technique combinations [82].

Ensemble Models Implementations

I have tried to implement various types of ensemble models by leveraging the top-performing models identified through comparative performance analysis. The following subsections detail the methodologies employed in the implementation.

Ensemble Model 1 This is a homogeneous ensemble model comprising linear regression variants, namely Ridge Regression (RidgeCV), Huber Regression (HuberRegressor), and Lasso Regression with Least Angle Regression (LassoLarsIC), is implemented using the scikit-learn library. The ensemble model is designed to leverage the predictive capabilities of each constituent linear model to enhance overall performance.

The constituent models are used as the base models. This ensemble model was constructed by stacking these base models, while the linear regression model was used as the final estimator. Each constituent model is optimized using hyperparameter tuning via GridSearchCV, aiming to enhance predictive accuracy.

Ensemble Model 2 This is a heterogeneous ensemble model comprising three distinct types of regression models: Multi-layer Perceptron Regressor (MLPRegressor), Partial Least Squares Regression (PLSRegression), and Support Vector Regressor (SVR). These models differ not only in their algorithms but also in their underlying mathematical principles and

assumptions. For example, MLPRegressor is a neural network-based model, PLSRegression is a linear regression technique, and SVR is a support vector machine-based model for regression tasks.

This model is implemented in a way that it trains the individual base regression models, MLPRegressor, PLSRegression, and SVR. Then a StackingRegressor ensemble model is constructed using these base models and a final estimator, GradientBoostingRegressor to make predictions. Again, it is implemented using the scikit-learn library.

The SVR model was constructed with hyperparameter tuning using GridSearchCV to find the best values for the regularization parameter and the epsilon-insensitive loss parameter.

Ensemble Model 3 For the third type of ensemble model, I stacked a Pytorch Temporal Convolutional Network (TCN) with the Multi-Layer Perceptron (MLP) and Partial Least Squares Regression (PLSR). This configuration achieved the lowest Root Mean Square Error (RMSE) score for Participant 1's 30-minute prediction horizon (PH). Plus, the stacked MLP and PLSR resulted in the lowest RMSE score for Participant 1's 60-minute prediction horizon, as well as for Participant 2's 30-minute and 60-minute prediction horizons. Lastly, the Ridge model was included as the base model because it had the best overall performance based on RMSE, MAE, Clarke Error Grid, and Parkes Error Grid for Participant 1's data. While the model with the best performance for Participant 2 was the same one with the lowest RMSE, which is the Stacked MLP and PLSR model and already included as the base model. When I experimented without using Ridge as the base model, the results were slightly worse for Participant 2's predictions on both PH30 and PH60, and for Participant 1's prediction on PH60. Although, there was a slight improvement in Participant 1's prediction on PH30. Thus, I decided to go with the one with Ridge because it showed more overall improvement.

The four base models, Ridge, MLP, PLSR, and TCN were trained independently. Afterward, the predictions from each base model are combined, and the mean of these predictions is used as the final prediction.

4

Result

This chapter consists of three distinct sections: a comparison study, multiple ML models' prediction analysis on data collected from two participants, and results obtained from using various approaches to improve predictions during physical activity (PA). The comparison study is divided into three subparts, each focusing on a different validation stage of the research. The analysis of participant-collected data provides an overview of model performance, including performance during PA and prediction after PA. Finally, the section on improving predictions during PA explores two approaches, a physiological hybrid method and the use of ensemble models.

4.1 Comparison Work

This section compares the prediction performance of models from five different papers with publicly available source code. The work is divided into three parts: validation of the models using the same datasets reported in the papers, validation with an Apple Health dataset collected from a participant, and finally, a performance comparison after implementing the models in GluPredKit.

4.1.1 Part 1 - Result of Validation

This section explores the performance of models for predicting glucose levels as reported in five research papers. These papers cover a range of prediction horizons, experimental configurations, and model approaches. The result is derived from using the same OhioT1DM dataset.

Paper 1 originally reported the Bidirectional LSTM model's performance with an RMSE of 37.40 mg/dL for a prediction horizon of 60 minutes, specifically trained using the Physical Activity (PA) feature. Upon running the model, the obtained result showed an average root mean squared error (RMSE) of 38.85 mg/dL for the same prediction horizon (60 minutes). It's noteworthy that in the nature of the model's operation, the PA feature was selected only once for patient 567, specifically for a Prediction Horizon of 60 minutes.

Paper 2 reported the results for both Prediction Horizon (PH) 60 and PH 30 minutes across different approaches (Basic and Stacked) and patients (540, 544, 552, 567, 584, 596) in terms of RMSE for Partial Least Squares regression (PLSR), Multilayer perceptron (MLP), and Long short-term memory (LSTM). What was observed is that the differences between the paper-reported and actual results are relatively minor.

For PH 60, the average difference between the paper-reported and actual RMSE across all approaches and patients ranges between 0.001 to 0.98 mg/dL, indicating a generally close alignment between the reported and actual values. Similarly, for PH 30, the average difference ranges from 0.07 to 0.70 mg/dL, again reflecting a relatively small deviation between the reported and actual results.

This observation suggests that while there are slight variations between the paper-reported and actual results, the differences are not substantial. In the context of glucose monitoring, this can be explained by the typical display conventions used for these devices. Glucose monitors often represent measurements in milligrams per deciliter (mg/dL) as integers and in millimoles per liter (mmol/L) with one decimal place. Given this level of precision, it can be said that small variations in reported and actual results can occur without affecting the overall interpretation of data.

The average difference between the paper-reported and actual root mean square error across all approaches and patients ranged from 0.001 to 0.98 mg/dL. This indicates a generally close alignment between the reported and actual values, suggesting high consistency. Similarly, with a Physical Health 30 approach, the average difference ranged from 0.07 to 0.70 mg/dL, also reflecting a relatively small deviation between reported and actual results.

This close alignment, particularly given the inherent rounding and measurement variability in glucose monitoring, supports the notion that the results presented in the paper are

reliable and accurate representations of the actual data collected.

The actual results consistently support the claims of the paper, which is the stacked regression technique significantly improved prediction performance, especially with a 30-minute history compared to a 60-minute history for Method 1. In addition, a similar pattern emerged in which the predictive performance of Method 2 closely matched that of Method 1 for most patients, with the exception of patient 552, which showed significantly worse performance. These results are consistent with the conclusion of the paper, suggesting that Method 1, which integrates average activity data into CGM windows, slightly outperforms Method 2, which trains models separately on CGM and activity data. Overall, Method 1, especially with a 30-minute history, showed superior performance.

Paper 3 has reported the results in varied experimental configurations. The performance comparison analysis based on the experimental configurations is presented in Table 4.1.

Table 4.1: Performance Comparison

Whatif	PH	Dropout	History	Features	RMSE - Actual(mg/dL)	RMSE - Paper (mg/dL)
No	60	0	30	BG	32.75	32.04
No	60	0	30	BG, I, M	31.80	30.94
No	30	0.1	30	BG, I, M	20.37	18.74
Yes	30	0.1	60	BG, I, M	19.40	18.19
Yes	60	0.1	60	BG, I, M	30.07	29.12
No	60	0.1	30	BG, I, M, SC, HR, ST	31.79	30.4
No	30	0.1	60	BG, I, M, SC, HR, ST	19.97	18.76

The table provides an overview of the different experimental configurations evaluated. Notably, it demonstrates the impact of varying parameters such as *Whatif*, *PH* (Prediction Horizon), *Dropout*, *History*, and *Features* on the Root Mean Squared Error (RMSE) values.

From the results, it is observed that while most configurations show comparable performance compared to the paper's reported RMSE values, the overall actual performance is slightly worse than the reported performance. Upon comparing the actual and paper-reported RMSE values for the configurations, the average difference was approximately 1.02(mg/dL). The range of differences varied up to 1.81(mg/dL) between the two sets of results. This difference may be attributed to unreplicated hyperparameters or specific settings not fully detailed in the paper or the available documentation.

Overall, these findings suggest the sensitivity of the model's performance to various input configurations, indicating potential enhancements under specific settings.

Paper 4 reported Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for both Multi-Task Learning (MTL) and single-task learning (STL) models at different prediction horizons. The MTL approach showcases slightly better RMSE (19.79 mg/dL vs. 20.67 mg/dL for 30 minutes; 33.73 mg/dL vs. 34.40mg/dL for 60 minutes) and MAE (13.83 mg/dL vs. 14.28 mg/dL for 30 minutes) compared to STL at the 30-minute horizon. However, at the 60-minute horizon, STL and MTL perform comparably, having almost identical RMSE (34.32 mg/dL) and MAE (24.97 mg/dL).

The validation process confirms the reported results for MTL at both 30 and 60 minutes. However, for STL, while the 60-minute RMSE and MAE match the reported values, there's a slight difference in the 30-minute RMSE and MAE, where the RMSE is slightly higher (20.67 mg/dL ± 0.32 compared to 20.67 mg/dL) and the MAE is slightly lower (14.28 mg/dL ± 0.19 compared to 14.28 mg/dL).

Paper 5 provided CGM-only RMSE values of 19.50 mg/dL and 34.36 mg/dL for prediction horizons (PH) of 30 and 60 minutes, respectively. For the NN-EIM (Error Imputation Module), the reported RMSE was 18.63 and 32.37 for PH 30 and 60 minutes. Echoing these findings, the validation process indicated CGM-only RMSE values of 19.22 mg/dL and 33.52 mg/dL, and for NN-EIM using selected features (CGM readings, CGM slope, and IOB), 18.78 and 32.15 mg/dL for PH 30 and 60 minutes.

The shallow neural network is trained with present and past CGM readings, CGM slopes, and IOB. The Error Imputation Module (NN-EIM) then creates a new feature pool termed Corrective Feature Set, encompassing first-order differences at various time lags of CGM, IOB, COB, sleep/work periods, skin temperature, and acceleration data. However, the paper does not explicitly detail the variable names, and the absence of guidance in the source code, such as a 'readme' file, hindered replication to achieve the same results.

Given the ambiguity in precisely identifying the features used in the reported NN-EIM, additional experimentation was conducted i.e. tested with various feature sets. Surprisingly, despite the inclusion of more features, the RMSE improvement was negligible, suggesting that the added features did not significantly enhance predictive performance.

4.1.2 Part 2 - Result of Applying Different Data

For this validation work, I made direct adjustments and manipulations to the code to assess the model's consistency across a different dataset. I received the Apple dataset provided by a participant, which spans over a year. However, to ensure comparability with the Ohio1 dataset, I reduced the data to three months. For simplicity, the results in Table 4.2 show the comparison of the average RMSE in the Ohio dataset with the RMSE from a single

participant's data in the Apple Health dataset, which is based on one participant. The result is based on the 60-minute prediction horizon.

Paper	Average RMSE (mg/dL) (Reported)	RMSE (mg/dL) (With Different Data)
Paper 1	37.40	38.39
Paper 2	30.79(Method 1), 30.71(Method 2)	30.92(Method 1), 31.13(Method 2)
Paper 3	31.6	31.45
Paper 4	STL: 34.40 MTL: 33.73	STL: 60.65 MTL: 62.04
Paper 5	31.4785	NA

Table 4.2: Summary of second validation work

Table 4.2 shows that the RMSE with a different dataset also generated a similar result for Paper 1 to Paper 3, while, applying a different dataset to the model from Paper 4 failed to produce the expected results. It should be noted that the multiple patients' data in the Ohio dataset is used to train the original work of paper 4. For paper 5, producing results using the Apple Health dataset has failed. Due to the time constraint of the project, I have moved on to the next comparison work.

4.1.3 Part 3 - Result of Implementing Models in GluPredKit

As a subsequent step, I integrated the models described in the papers into the GluPredKit platform. This is conducted as a progress towards integrating the platform into my thesis and for the better scalability of handling different datasets for different models.

Table 4.3: RMSE(mg/dL) of the models in each paper after integrating into GluPredKit, with the differences from reported values shown in parentheses

Paper	**30PH**	**60PH**
Paper1	28.21 (+8.01)	39.51 (+5.32)
Paper2	19.96 (+0.77)	32.59 (-1.92)
Paper3	17.42 (-1.58)	32.88 (1.94)
Paper4	19.36[STL] (+1.31), 20.30[MTL] (-0.51)	35.92[STL] (-1.52), 32.92[MTL] (+0.81)
Paper5	91.78 (+73.15)	90.87 (+58.60)

Table 4.3 displays the RMSE (in mg/dL) of models from the five papers after integrating them into GluPredKit, focusing on data from a single patient (570 of Ohio). This approach is taken just to facilitate a clear and simple representation of the implementation results. The values in parentheses represent the difference between the new results and the reported results. Positive differences indicate that the new results are higher than the reported ones, while negative differences suggest they are lower. This information can be used to evaluate the impact of integrating these models into GluPredKit.

For papers 2 to 3, the result has shown a similar outcome, while the model from papers 1 and 4 generated slightly worse performance than the reported. Notably, the last model, when integrated, exhibited significantly poorer performance, signaling a failure in its integration within the framework. Translating the original code which was written in Matlab and some Matlab-specific libraries, to Python was a challenge.

4.2 Work with Participant Collected Data

With data collected by two participants, the models were compared based on their overall performance. Not just during the participants' physical activity periods, the total prediction accuracy over the entire test dataset was evaluated. The metrics used for comparison included RMSE, MAE, Clarke Error Grid, and Parkes Error Grid. The specific prediction models are discussed in the "Evaluated Models" section of the "Design" chapter.

4.2.1 Overall Performance Comparison

The performance of each model is compared using RMSE, MAE, Clarke Error Grid, and Parkes Error Grid. The Clarke and Parkes Error Grids were evaluated based on the proportion of prediction points that fell into zone A—the higher the proportion, the better the model's score. The models were tested for prediction horizons of 30 minutes and 60 minutes.

Intriguingly, the optimal model varied across datasets. For Participant 1 data, TCN implemented in PyTorch emerged as the top performer in terms of both RMSE and MAE, with the Stacked MLP and PLSR model as the second-best performing model for both 30 and 60-minute prediction horizons. Notably, the TCN model achieved the best performance in the Clarke Error Grid for the 30-minute prediction horizon, while the Gradient-boosted trees (GBT) model stood out for the 60-minute horizon. The ARX model exhibited superior performance in the Parkes Error Grid for both prediction horizons for Participant 1's data. Conversely, for Participant 2 data, Stacked MLP and PLSR demonstrated the best RMSE and MAE performance, alongside the Parkes Error Grid for both prediction horizons. However, the Clarke Error Grid highlighted Support Vector Regression (SVR) Linear for the 30-minute horizon and LSTM for the 60-minute horizon as the best-performing models.

Overall Performance Score

Across all metrics, Ridge, PLSR, ARX, Stacked MLP and PLSR, Temporal Convolutional Networks (TCN), GBT, and Random Forest consistently ranked within the top 5 for Participant 1 data. For Participant 1's data, Ridge, PLSR, and ARX claimed the top three positions for

the overall score. This overall score was arbitrarily calculated by assigning the highest score to the top-performing model, the lowest score to the fifth-ranked model, and then summing across all metrics, RMSE, MAE, Clarke Error Grid, and Parkes Error Grid. In contrast, for Participant 2 data, Stacked MLP and PLSR, Huber, SVR (radial basis function), ARX, Ridge, SVR Linear, GBT, and LSTM emerged as the top performers across various evaluation metrics. Notably, Stacked MLP and PLSR, Huber, and SVR (radial basis function) secured the top 3 positions for overall score calculation.

Clarke Error Grid

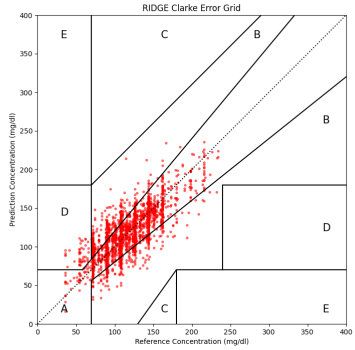

Figure 4.1: Clarke Error Grid for Participant 1 with a Ridge model (30-minute prediction horizon)

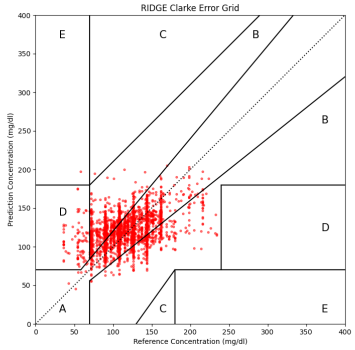

Figure 4.2: Clarke Error Grid for Participant 1 with a Ridge model (60-minute prediction horizon)

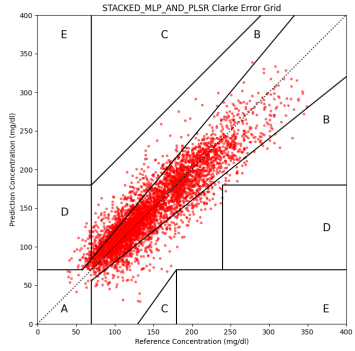

Figure 4.3: Clarke Error Grid for Participant 2 with a Stacked model (MLP and PLSR) (30-minute prediction horizon)

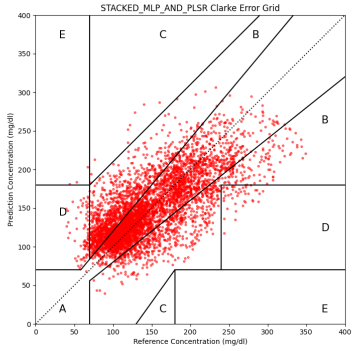

Figure 4.4: Clarke Error Grid for Participant 2 with a Stacked model (MLP and PLSR) (60-minute prediction horizon)

The Clarke Error Grids for the participants' data depict the performance of the best-performing models, identified by their overall performance scores. It is evident that as the prediction horizon extends from 30 minutes to 60 minutes, the prediction accuracy declines, resulting in some predictions falling within the clinically significant zones (C, D, E), indicating potential risks.

Although both participants had similar amounts of data (3 months of data logged every 5 minutes), there seem to be more data points for Participant 1 in the Clarke Error Grid plot. This could be because Participant 1's blood glucose (BG) levels stayed within a narrower range, while Participant 2's BG levels fluctuated over a wider range.

Parkes Error Grid

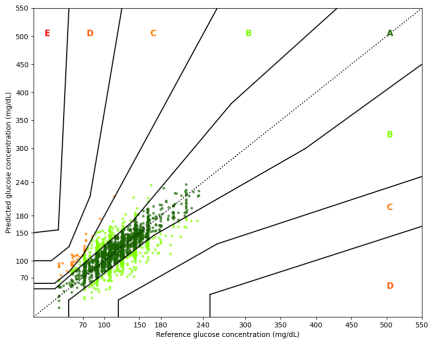

Figure 4.5: Parkes Error Grid for Participant 1 with a Ridge model (30-minute prediction horizon)

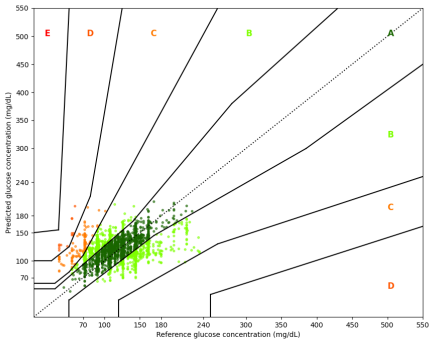

Figure 4.6: Parkes Error Grid for Participant 1 with a Ridge model (60-minute prediction horizon)

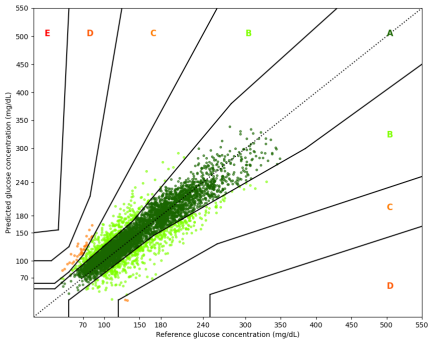

Figure 4.7: Parkes Error Grid for Participant 2 with a Stacked model (MLP and PLSR) (30-minute prediction horizon)

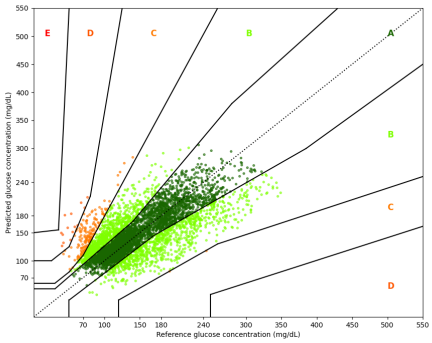

Figure 4.8: Parkes Error Grid for Participant 2 with a Stacked model (MLP and PLSR) (60-minute prediction horizon)

The above figures show the Parkes Error Grids for the participants' data depict the performance of the best-performing models, identified by their overall performance scores. The same trend is seen here. As the prediction horizon increases, more prediction points end up in the dangerous zones (C, D, E). However, an interesting finding is that the same model predictions can fall into different zones in the Parkes Error Grid compared to the Clarke Error Grid. For example, there are predictions no longer fall into Zone E in the Parkes Error Grid, while there were some predictions in Zone E in the Clarke Error Grid.

It seems that Zone E in the Parkes Error Grid has different thresholds for hypoglycemia or hyperglycemia compared to Zone E in the Clarke Error Grid. In fact, the Parkes error grid

is simpler than the Clarke error grid. In the Parkes grid, the zones start from the center and spread out to the edges, while the Clarke grid has sharp lines, and some zones are skipped. Even if a blood glucose meter is quite accurate, it might not meet the criteria for Zone A in the Clarke grid and may end up in Zone D instead. This may be caused by the fact that the Clarke grid was made for education, not as a strict accuracy standard. It is expected that the Parkes grid fixed some problems by making the zone boundaries continuous [89].

4.2.2 Performance During the Physical Activity

This section explores the performance evaluation result during physical activity (PA) across different models and participants. The analysis includes bar charts, box plots, and trajectory plots to illustrate how model performance varies during PA periods.

RMSE during Physical Activity Comparison

The RMSE values during the physical activity (PA) period did not display a consistent pattern, either significantly higher or lower than the overall RMSE. The results showed no clear trend.

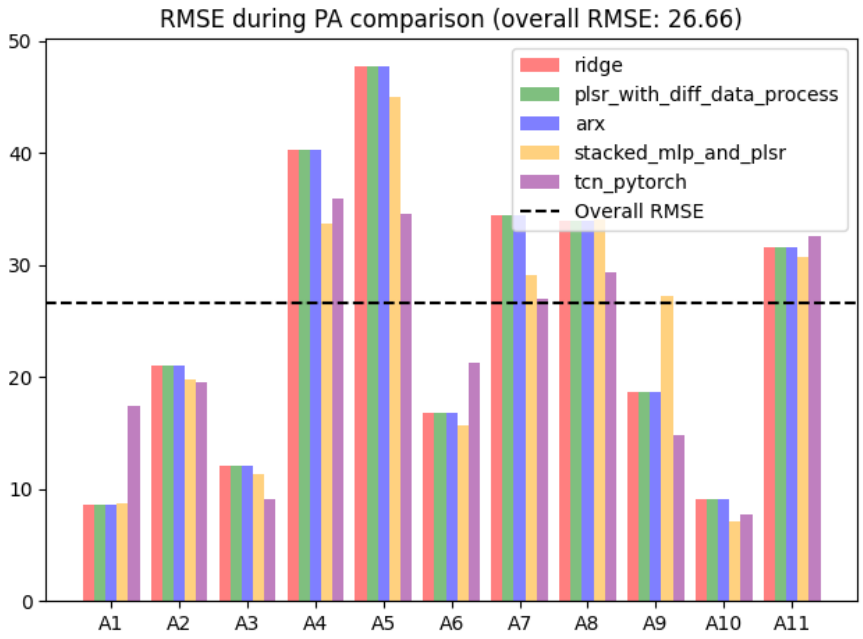

Figure 4.9: Top 5 models RMSE during PA for Participant 1

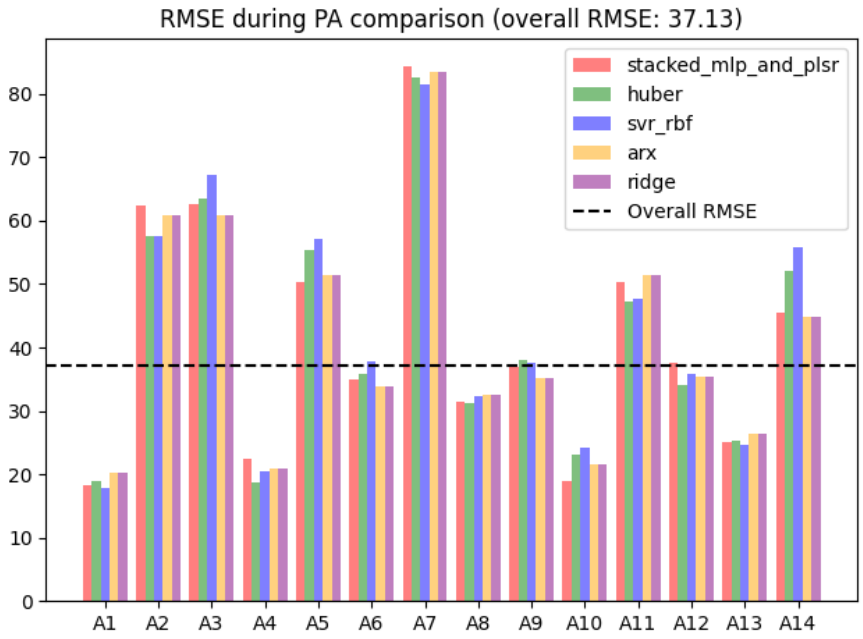

Figure 4.10: Top 5 models RMSE during PA for Participant 2

Neither Figure 4.9 nor Figure 4.10 shows RMSE values that consistently remained above or below the overall RMSE value, which is represented by the black dotted line in the figures.

Box Plot during PA

The box plot illustrates RMSE values during physical activities. The orange line represents the mean RMSEpa (RMSE during PA), while the lower and upper bars depict the variability of RMSE during PA. Additionally, the overall RMSE (RMSE during the entire test data) is indicated at the top of the plot.

Upon training with Participant 1's data, the TCN model (PyTorch implementation version) displayed the lowest overall RMSE score of 26.66 and the lowest RMSEpa (RMSE during physical activity) score of 22.66. This seems to reveal a pattern where the model with the lowest overall RMSE also achieves the lowest RMSE during physical activity. However, this pattern contrasts with the findings observed from Participant 2.

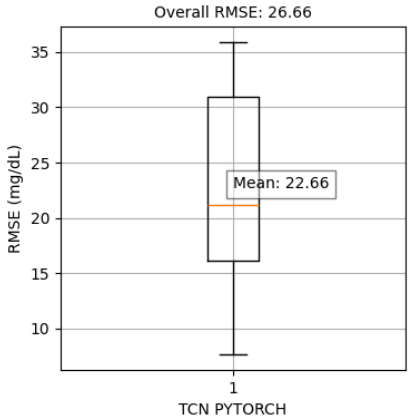

Figure 4.11: A box plot depicting the RMSE of TCN model with a 60-minute PH for Participant 1

For Participant 2, the LSTM model demonstrated the lowest RMSE score during the physical activity (PA) period among the models in comparison, indicating minimal variability, despite being ranked second worst in terms of overall RMSE values, following the TCN model.

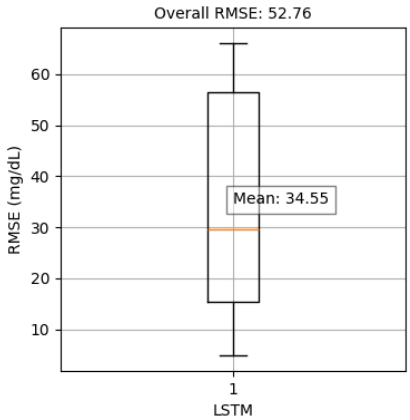

Figure 4.12: A box plot depicting the RMSE of LSTM model with a 60-minute PH for Participant 2

The mean RMSEpa (RMSE during PA) of the LSTM model, 34.55, is observed to be lower than that of the Stacked model (MLP and PLSR), which stands at 41.51. Despite the significantly lower overall RMSE of the Stacked model (36.70) compared to the LSTM (52.76), this contrast in mean RMSEpa values is deceiving.

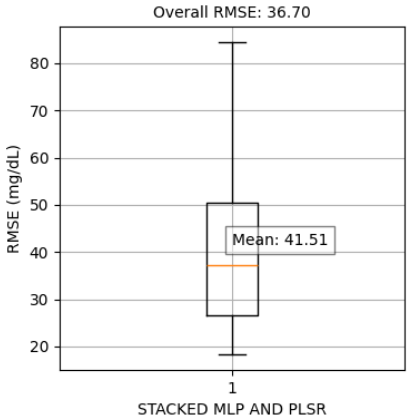

Figure 4.13: A box plot depicting the RMSE of Stacked model (MLP and PLSR) with a 60-minute PH for Participant 2

Upon analyzing the trajectory plot of the LSTM model as shown in Figure 4.14, it was observed that the predictions consistently converged toward the median value, indicating that the model is not trained well to pick up the dynamics of the blood glucose level using the trained features. This observation suggests that PA occurrences may have just luckily coincided with glucose levels close to the median value. It was also verified during follow-up interviews with participants, who reported exercising cautiously to avoid hypoglycemia episodes.

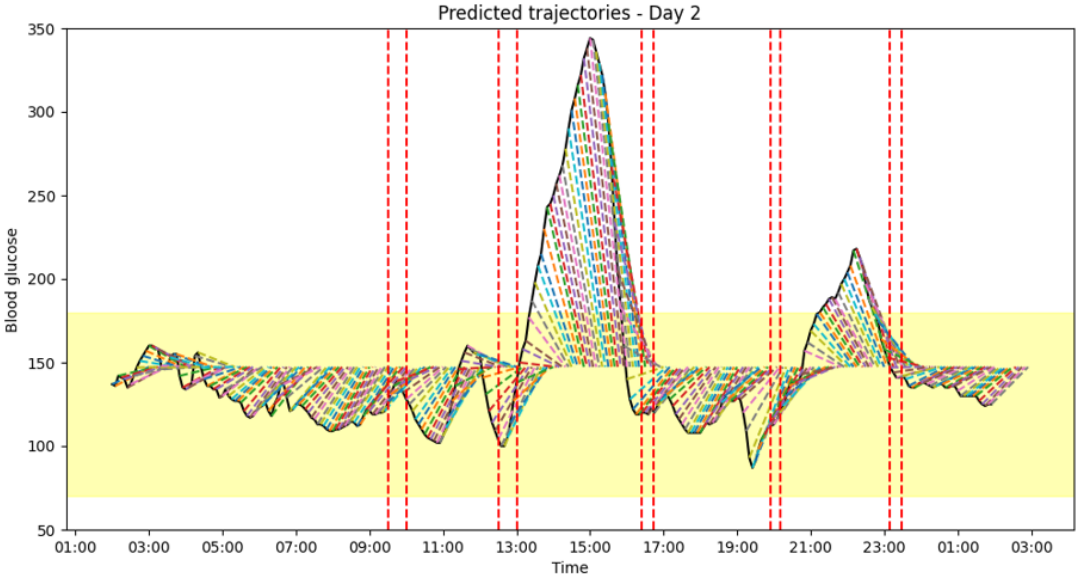

Figure 4.14: Trajectory plot of LSTM model with a 60-minute PH for Participant 2

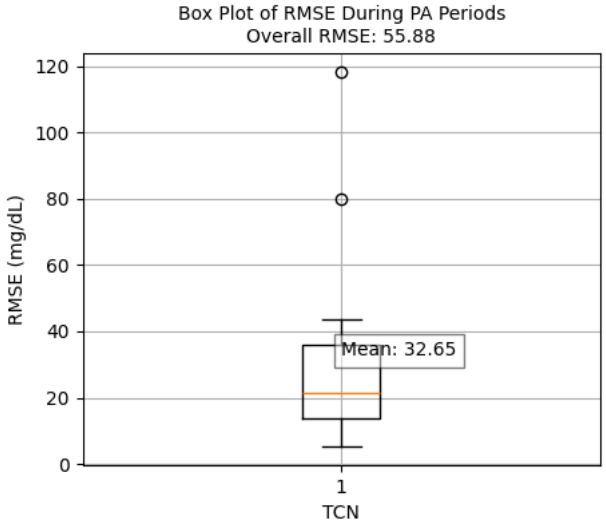

Figure 4.15: A box plot depicting the RMSE of TCN model with a 60-minute PH for Participant 2

Notably, Figure 4.15 shows that the TCN model on Participant 2's data exhibited significant variability, consistent with its higher overall RMSE value, as anticipated. Surprisingly, the TCN model achieved the lowest RMSEpa compared to other models, prompting further inquiry into whether RMSE effectively assesses predictive performance during physical activity in this study context.

Trajectory Plot per Day

This type of plot as Figure 4.16 and Figure 4.17, shows the complete trajectory over a day, with the physical activity (PA) periods highlighted by red dotted lines. Although the plot does not elucidate any noticeable changes in trajectory during PA, it does reveal the unique characteristics of the participant. Participant 1 consistently maintained their blood glucose levels within the safe range of 70 to 180 mg/dL, while Participant 2 exhibited more frequent excursions beyond this range, resulting in different predictive outcomes by the model. Examining individual trajectory patterns, Participant 1's trajectories tended to converge around a value of 130, whereas Participant 2's trajectories tended to stabilize around 150. This discrepancy may be attributed to Participant 2 consistently having higher blood glucose levels compared to Participant 1. Despite these differences, both examples illustrate the model's consistently conservative prediction of a decrease in blood glucose levels once they surpass approximately 170 mg/dL.

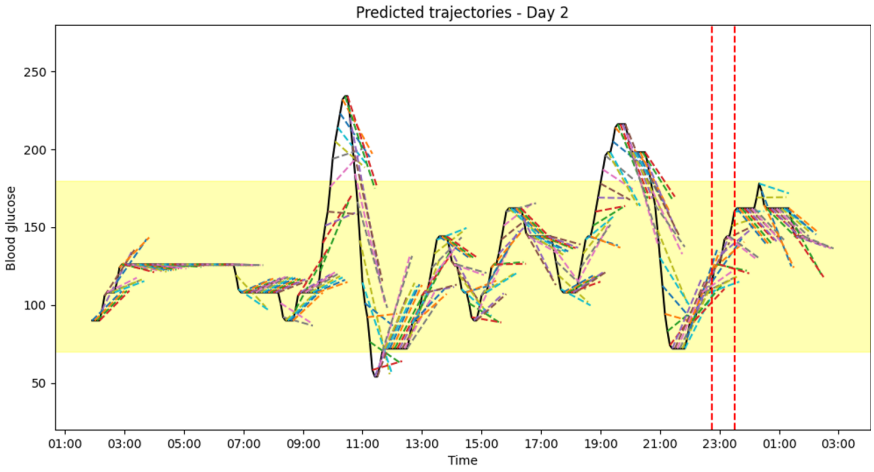

Figure 4.16: A Ridge model trajectory plot of Participant 1

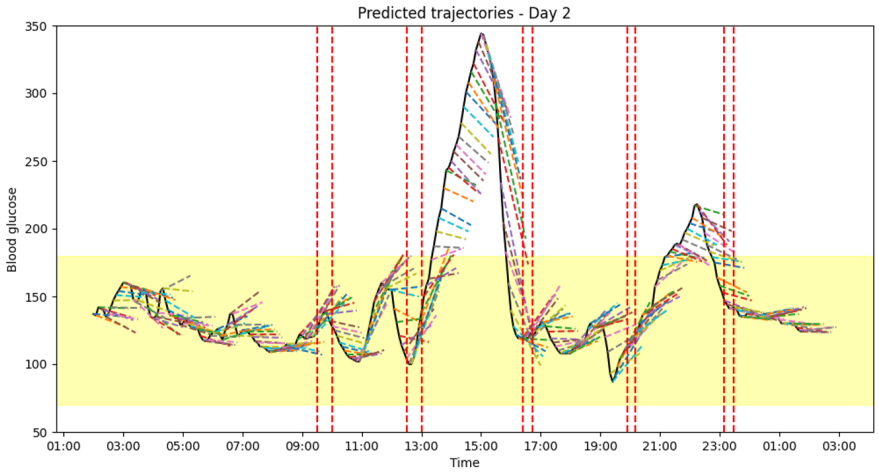

Figure 4.17: A Ridge model trajectory plot of Participant 2

The red lines indicate the periods during which physical activities were conducted. Upon reviewing these plots, I recognized the necessity of creating separate trajectory plots to more effectively illustrate the trajectories during physical activities. Consequently, I generated trajectory plots specifically focused on RMSE during physical activities.

Trajectory Plots with RMSE during PA

Trajectory plots with RMSE during the PA were analyzed to evaluate the actual predictions in detail for each model. While analyzing Recurrent Neural Network (RNN) models, specifically

Long Short-Term Memory (LSTM) and Temporal Convolutional Networks (TCN), during periods of physical activity (PA). These plots revealed important insights that RMSE alone cannot capture about the predictive behavior of these models.

What trajectories have revealed on RNN models For further analysis, trajectory plots were generated only for the PA periods, with the RMSE values calculated for each period. Upon examining these trajectory plots, I discovered an intriguing observation regarding the predictive capabilities of linear regression in comparison to LSTM and TCN models. While Long Short-Term Memory (LSTM) and Temporal Convolutional Networks (TCN) are widely recognized recurrent neural network architectures for time series analysis, particularly in predicting blood glucose levels, the trajectory plots revealed a tendency for these models to rapidly converge toward the mean value. Consequently, despite seemingly favorable RMSE performance scores, this behavior challenges the common assumption of LSTM and TCN models' efficacy in capturing the time-series characteristics of the data. These findings suggest that LSTM and TCN models may tend to converge toward the mean value quickly, resulting in lower RMSE scores that might falsely indicate good performance. This highlights again the importance of not relying solely on traditional evaluation metrics like RMSE and emphasizes the need for a deeper understanding of model behavior and performance.

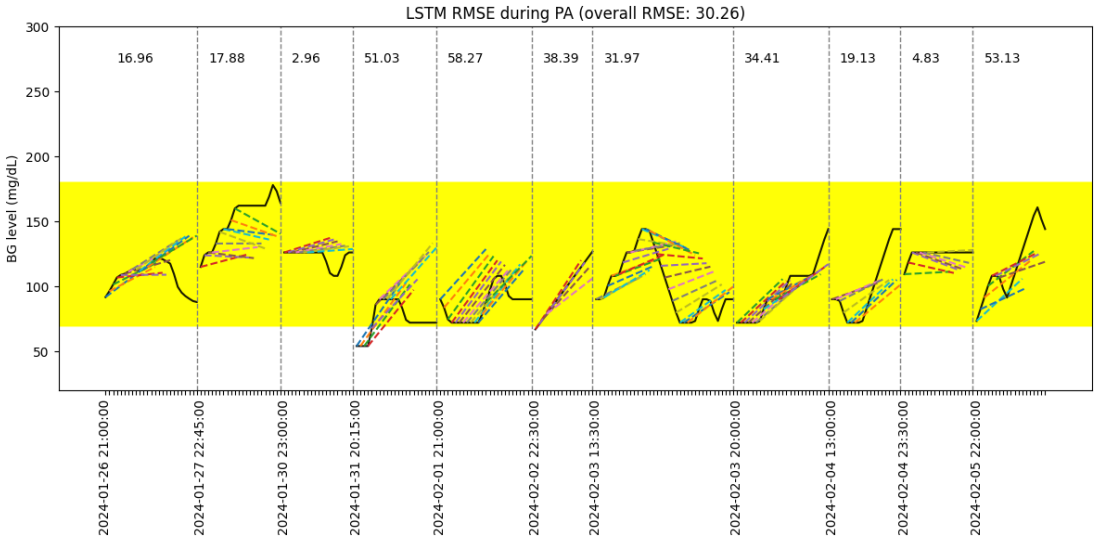

Figure 4.18: LSTM - Trajectories during PA for Participant 1

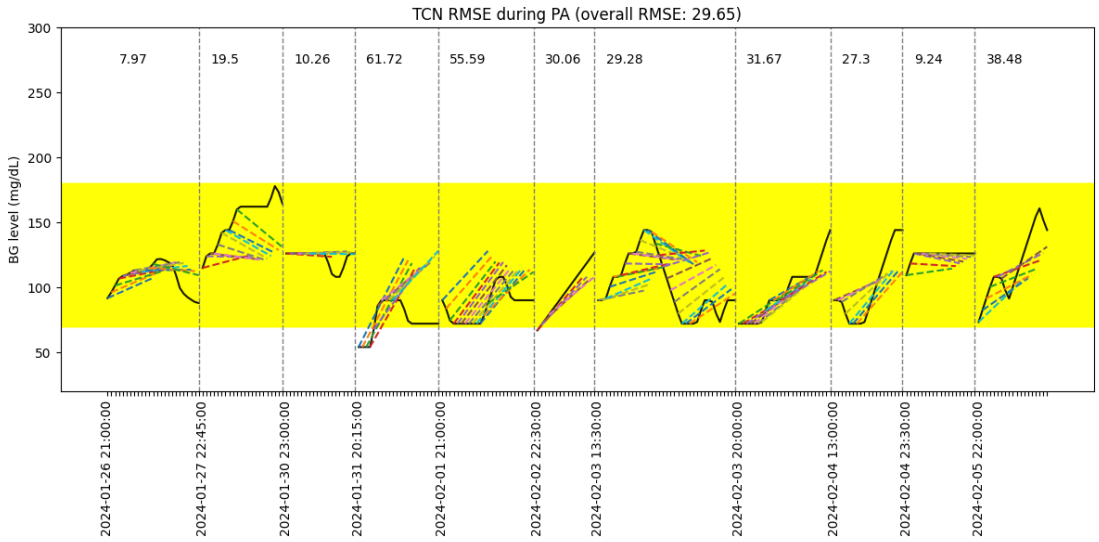

Figure 4.19: TCN - Trajectories during PA for Participant 1

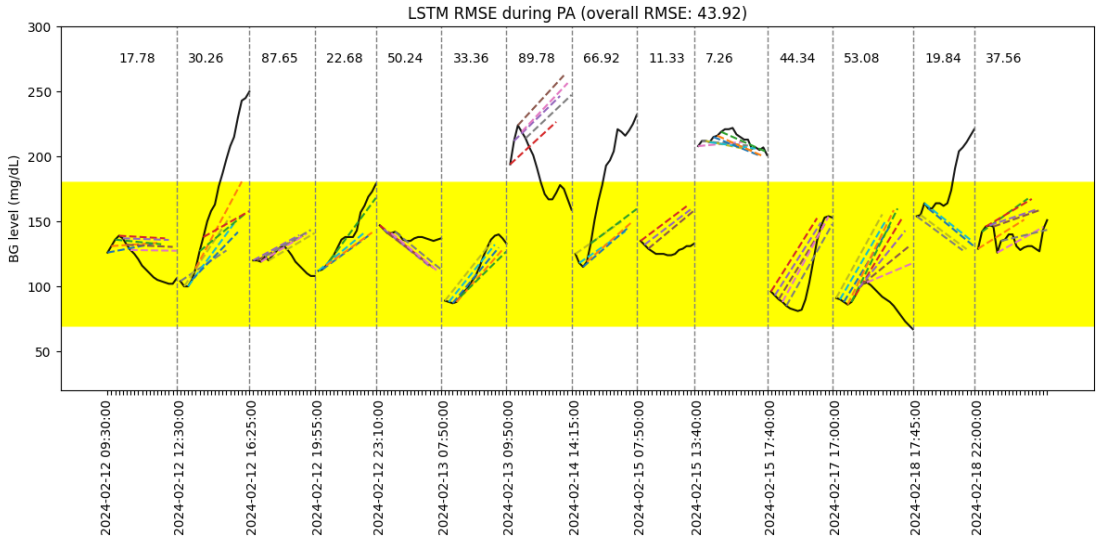

Figure 4.20: LSTM - Trajectories during PA for Participant 2

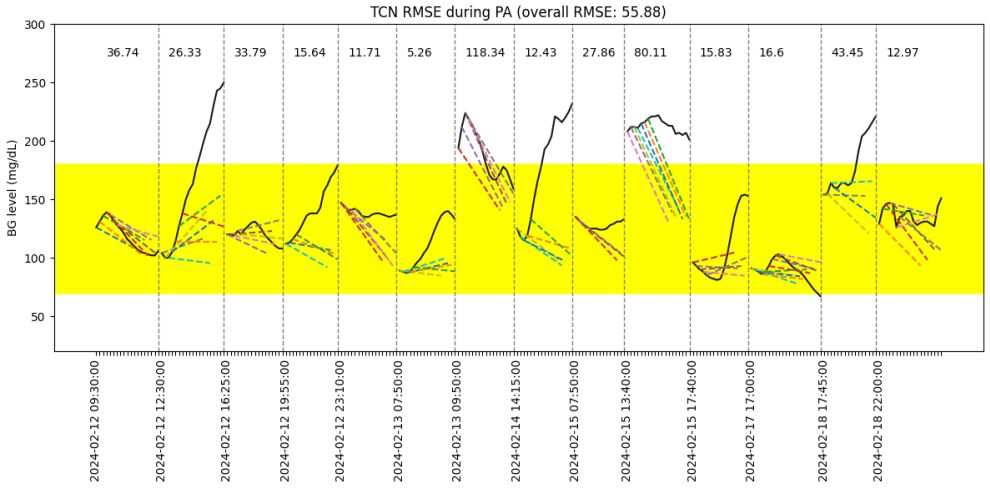

Figure 4.21: TCN - Trajectories during PA for Participant 2

Moreover, it's important to mention that this pattern seems to weaken when we have more training data and different types of model setups. For example, it can be observed that in Figure 4.22 and Figure 4.23, the TCN model in Pytorch implementation shows better trajectories and RMSE scores. This suggests that there are other things affecting how well the model works, not just having more data. This implies that it is important to understand the model's behavior in detail, including how it's built and designed. There could be other factors affecting how well these models perform, like adjusting certain settings, designing the structure of the model, and preparing the data before using it.

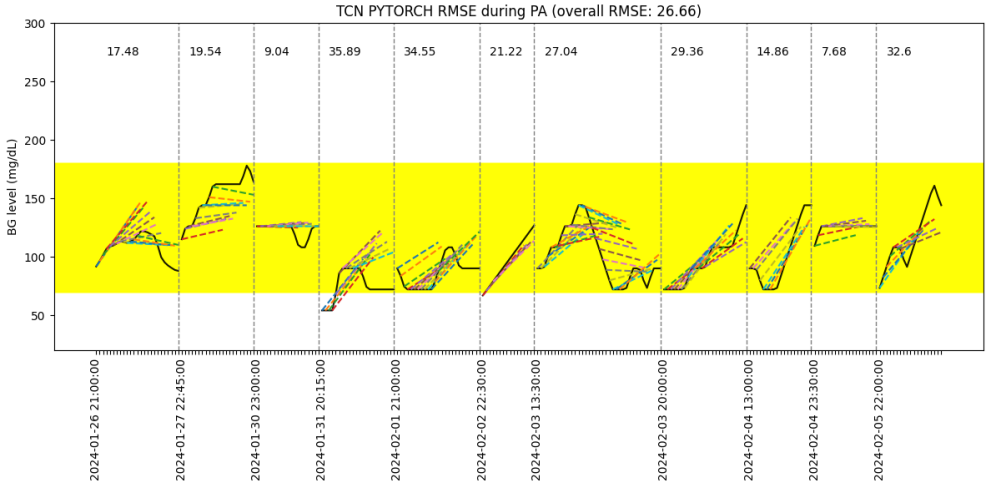

Figure 4.22: TCN Pytorch version - Trajectories during PA for Participant 1

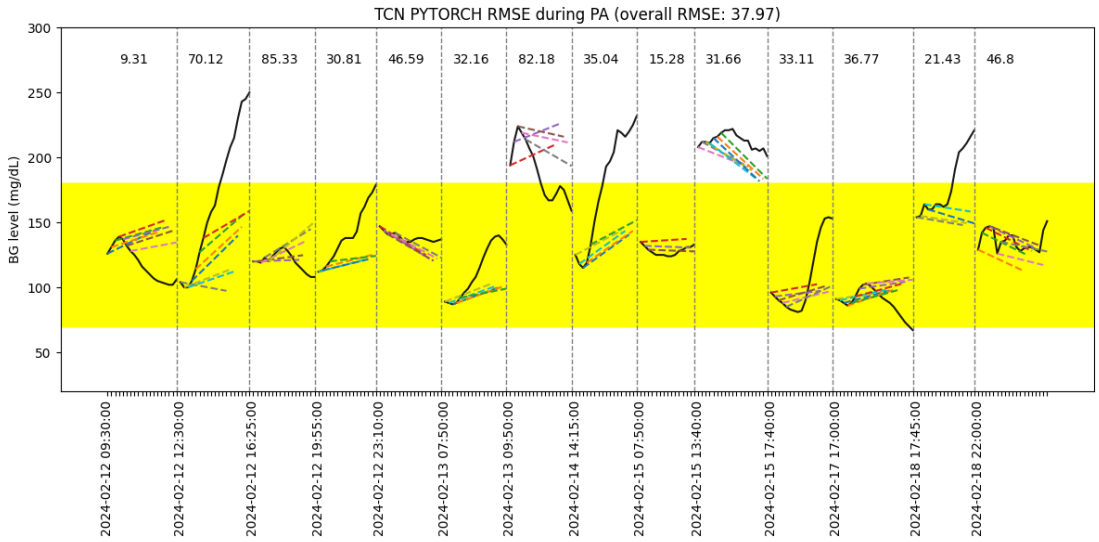

Figure 4.23: TCN Pytorch version - Trajectories during PA for Participant 2

Further investigation into what drives this behavior could provide valuable insights into how to better leverage LSTM and TCN models for time series prediction tasks. It's essential to approach model evaluation with caution, considering factors beyond RMSE scores, to ensure robust and reliable predictions in time series analysis.

What RMSE alone can't reveal The following two figures depict the trajectory plots with RMSE (Root Mean Square Error) values during the physical activity (PA) for a double LSTM model with a 60-minute prediction horizon for each participant. Figures 4.24 and 4.25 suggest that RMSE might not accurately represent the model's behavior. In Figure 4.24, the model tends to predict values close to the mean blood glucose level, resulting in a lower RMSE compared to Figure 4.25. The lower RMSE for Participant 1 can be attributed to their blood glucose levels staying closer to the mean value of the overall test data. Conversely, although the double LSTM model's predictions for Participant 2 demonstrate more varied behavior, with less tendency to predict the mean, they yield a higher RMSE than the predictions for Participant 1.

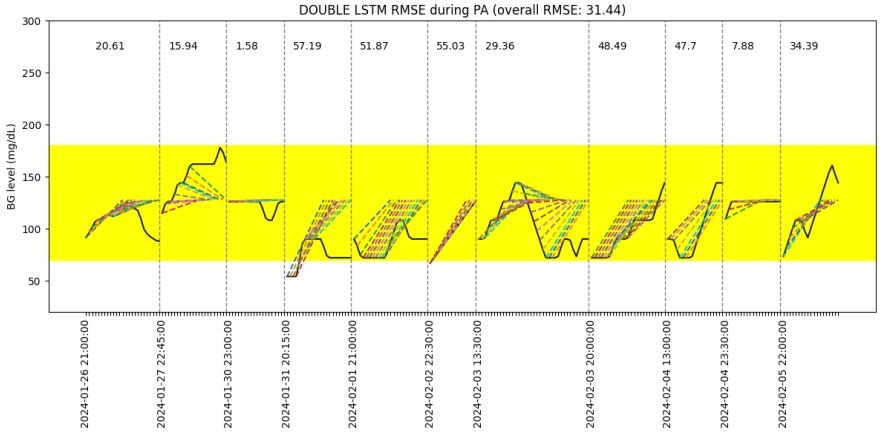

Figure 4.24: Double LSTM - Trajectories during PA for participant 1

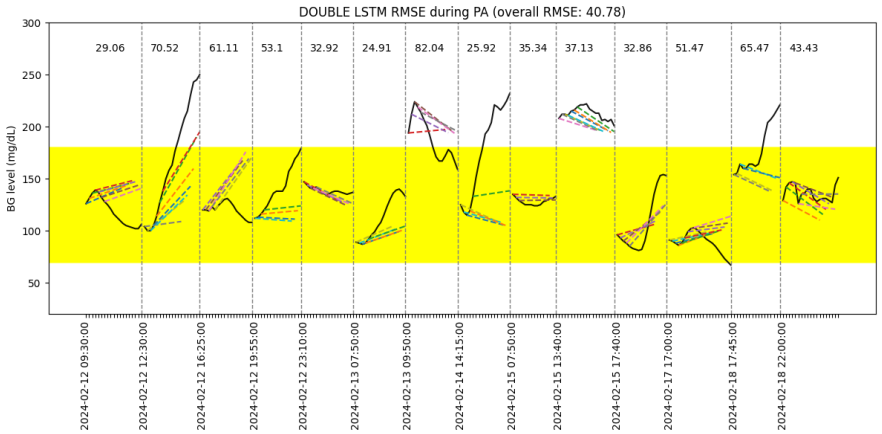

Figure 4.25: Double LSTM - Trajectories during PA for participant 2

4.2.3 Performance After Physical Activity

I examined whether there were significant changes in blood glucose levels after physical activity (PA) by comparing RMSE values and their trends at various time intervals. First, I compared RMSE values during different periods after the start of PA. Then I compared RMSE values at various durations following the end of PA.

After the start of Physical Activity

The following two figures illustrate how RMSE changes when calculated for different PA durations, ranging from 15 minutes to 1 hour after the start of physical activity.

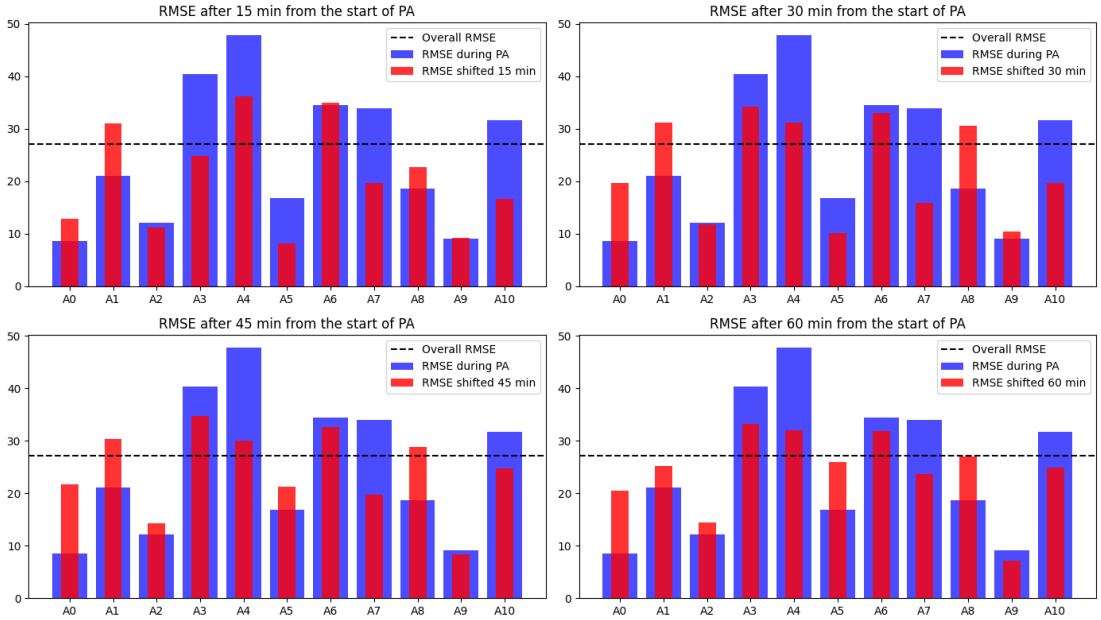

Figure 4.26: Bar chart showing RMSE for the Ridge model with a 60-minute prediction horizon after the start of physical activity for Participant 1

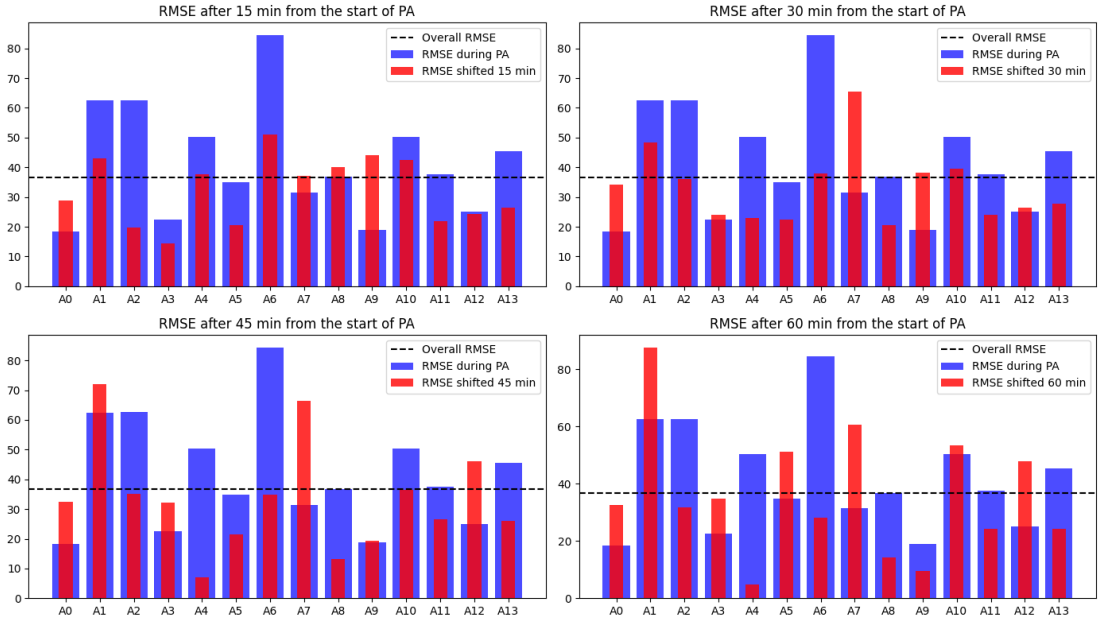

Figure 4.27: Bar chart showing RMSE for the Stacked MLP and PLSR model with a 60-minute prediction horizon after the start of physical activity for Participant 2

In the above figures, the best-performing model for each participant was selected based on the overall performance scores described in the 'Overall Performance Score' subsubsection. The blue bars show the RMSE value during the PA while the red bars represent the RMSE value after different periods from the start of PA. Figures 4.26 and 4.27 do not reveal any consistent patterns.

After the end of PA

I further investigated whether there were any noticeable changes in blood glucose levels after physical activity (PA) by comparing RMSE and trajectories at different time intervals, i.e. for 15 minutes, 30 minutes, 60 minutes, and 90 minutes each. In the following figures, the best-performing model for each participant was selected again to compare. Figure 4.28 and Figure 4.29 do not reveal any consistent patterns in trajectory lines.

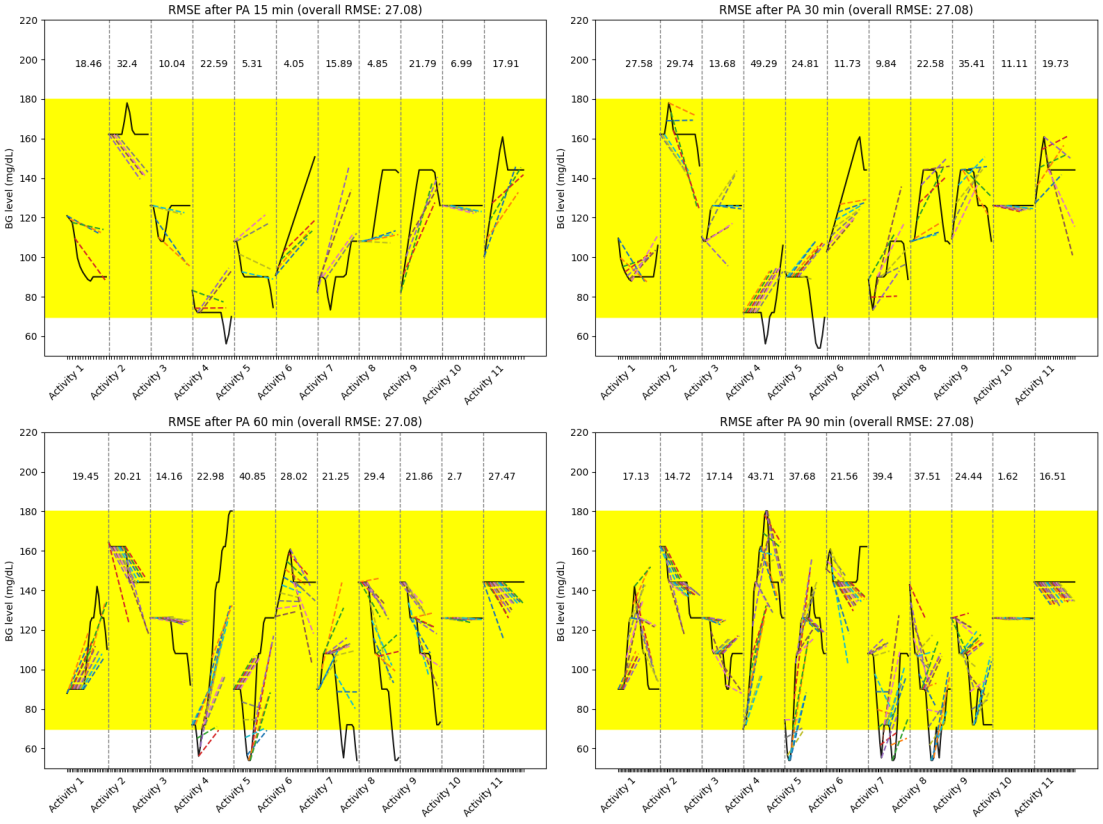

Figure 4.28: Four trajectory plots with RMSE for different time periods after PA for the Ridge model with 60-minute PH for Participant 1

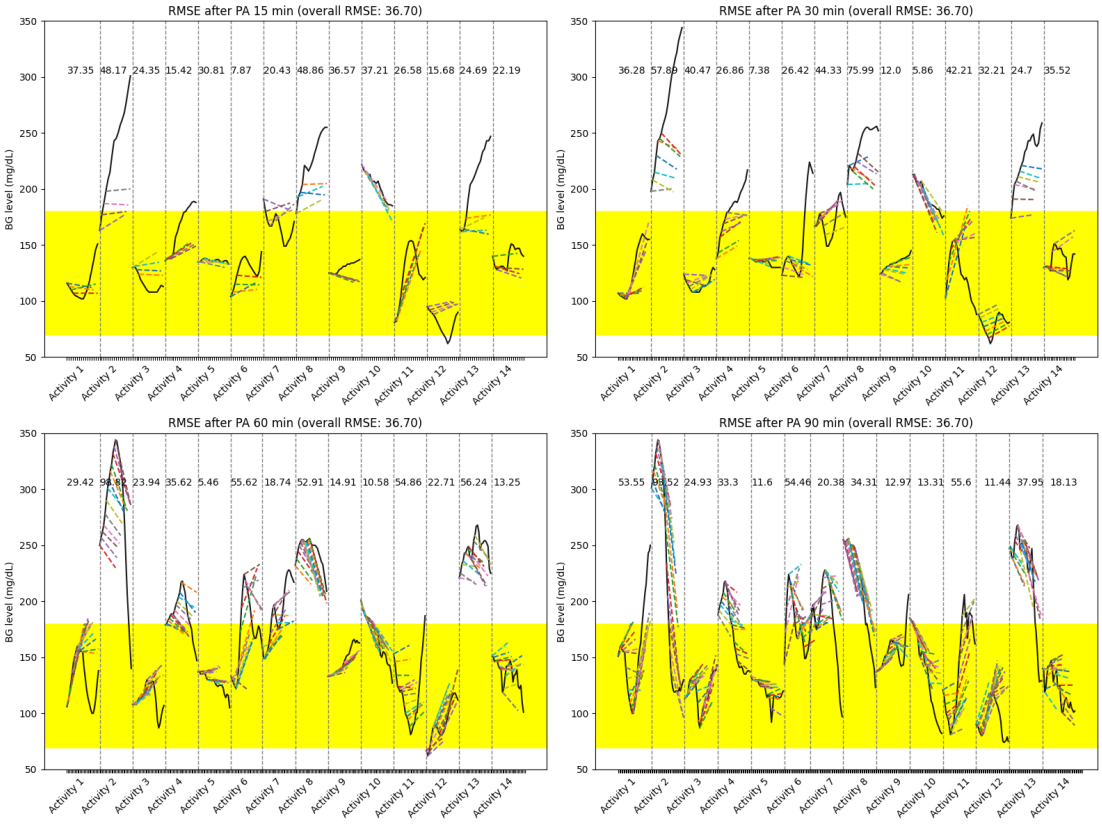

Figure 4.29: Four trajectory plots with RMSE for different time periods after PA for the Stacked MLP and PLSR model with 60-minute PH for Participant 2

I created a bar chart for further comparison in varying RMSE values. However, Figure 4.30 and Figure 4.31 did not reveal any consistent patterns either.

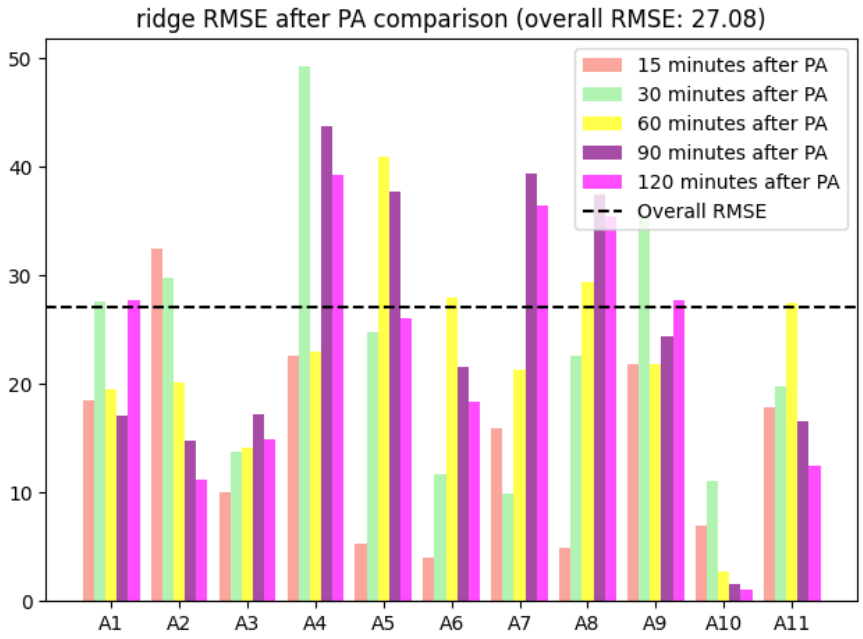

Figure 4.30: Four bar charts with RMSE for different time periods after PA for the Ridge model with 60-minute PH for Participant 1

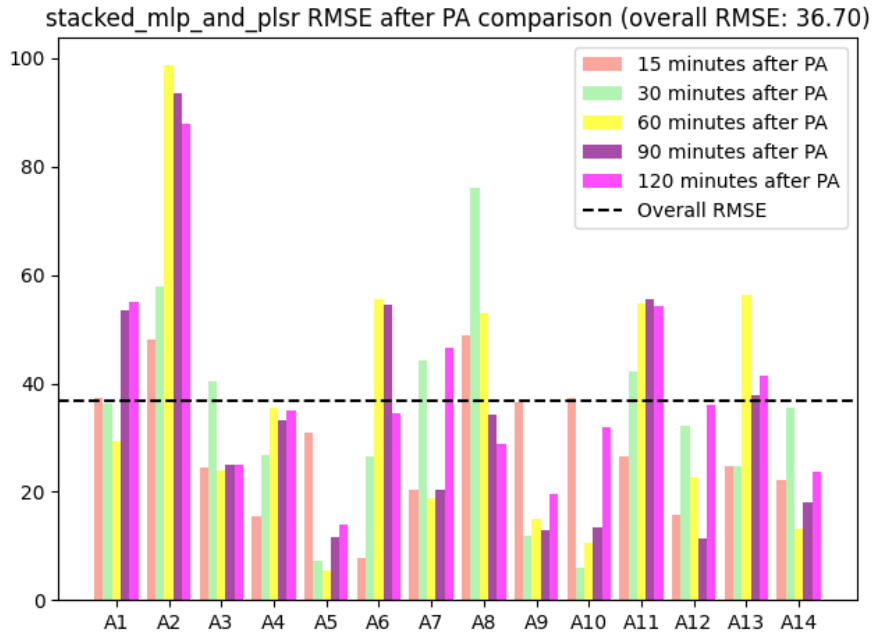

Figure 4.31: Four bar charts with RMSE for different time periods after PA for the Stacked MLP and PLSR model with 60-minute PH for Participant 2

In summary, the RMSE or trajectories don't consistently increase or decrease across different periods after PA.

4.3 Approaches Taken to Improve Predictions During Physical Activity

In this section, I explored the performance of several models I developed to improve blood glucose (BG) level predictions during physical activity. The models include a physiological hybrid model and three different ensemble models. The goal was to enhance prediction accuracy despite the dynamic and unpredictable nature of BG during physical activity.

4.3.1 Physiological Hybrid Model Performance During Physical Activity

The architecture of the physiological hybrid model, as detailed in the implementation section, is a prediction system that uses multiple Recurrent Neural Networks (RNNs) to analyze various types of input data and blood glucose readings. Each RNN is designed to process a specific type of input, learning patterns over time. The outputs from these different RNNs are then combined to create a prediction of future blood glucose levels.

Root Mean Squared Error

Figure 4.32 depicts the mean differences between the RMSE of physiological hybrid models' predictions and that of the benchmark model. The benchmark model was chosen based on the dataset for each participant. For Participant 1, the TCN implemented in PyTorch was selected because it had the lowest RMSE score, and its RMSE values and trajectory patterns were superior to those of the Ridge model, which otherwise had the best overall performance score. For Participant 2, the chosen benchmark was a stacked model combining an MLP and PLSR, as it achieved the best overall performance.

Here, the TCN-based model was trained with data from Participant 1, while the LSTM-based model was trained with data from Participant 2. These models share the same architectural structure as described in the implementation section. The results were compared with different physiological hybrid models because the LSTM-based model exhibited abnormal behavior when trained with Participant 1's data, as it tended to predict a constant mean value on the trajectory plot. Conversely, the TCN-based model performed worse than the LSTM-based model when trained with Participant 2's data.

To determine whether the RMSE during physical activity (RMSEpa) has improved, the RMSEpa of the Physiological Hybrid model is subtracted from that of the benchmark model. If the mean difference in RMSEpa is negative, it indicates that the Physiological Hybrid model has a lower average RMSE during physical activity compared to the benchmark.

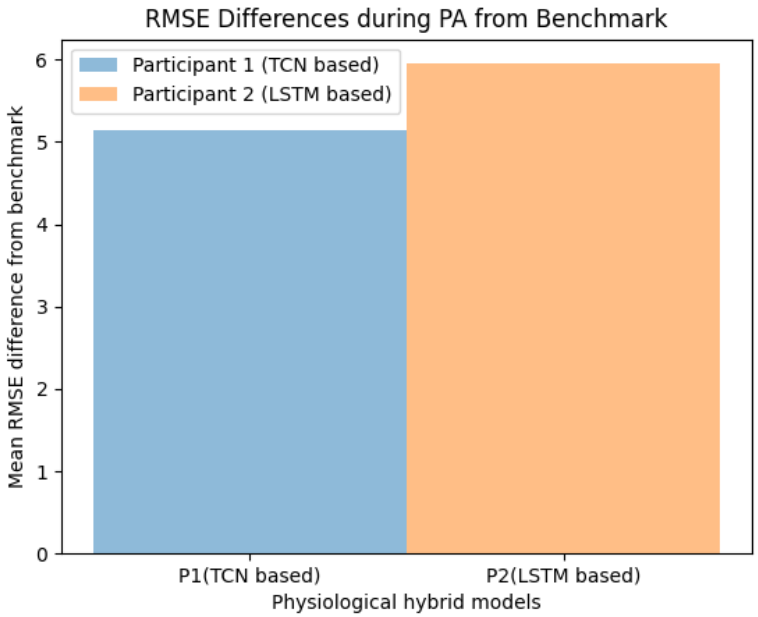

Figure 4.32: Bar chart on Mean RMSE Differences from Benchmark Models

The results in Figure 4.32 show that the Physiological Hybrid model did not enhance prediction during physical activity. Both models exhibited higher RMSEpa values compared to the benchmark models. It's worth noting, however, that while the overall RMSE of the TCN-based Physiological Hybrid model for participant 1 data (26.18) was slightly lower than that of the benchmark model (26.66), the RMSE during physical activity was worse. In contrast, the LSTM-based model for Participant 2 also had a higher overall RMSE score.

Trajectory Plot During PA

The following figures are the trajectory plots for each Physiological Hybrid Model with RMSE values for each PA period.

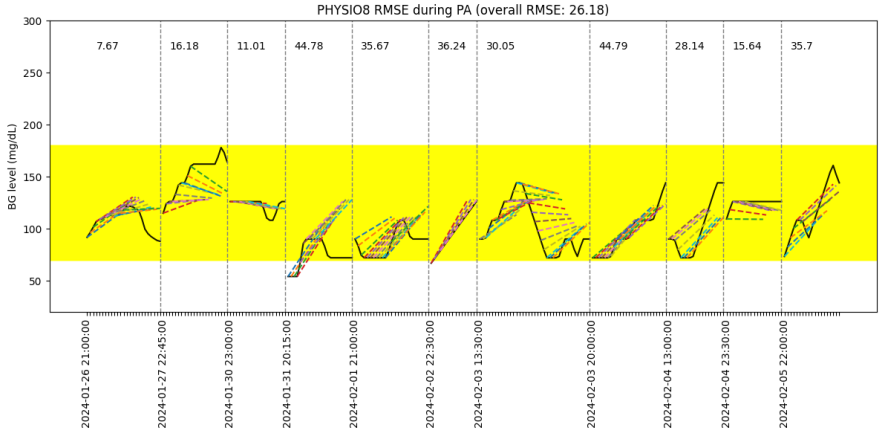

Figure 4.33: Trajectory plot on TCN-based Physiological Hybrid Model for Participant 1

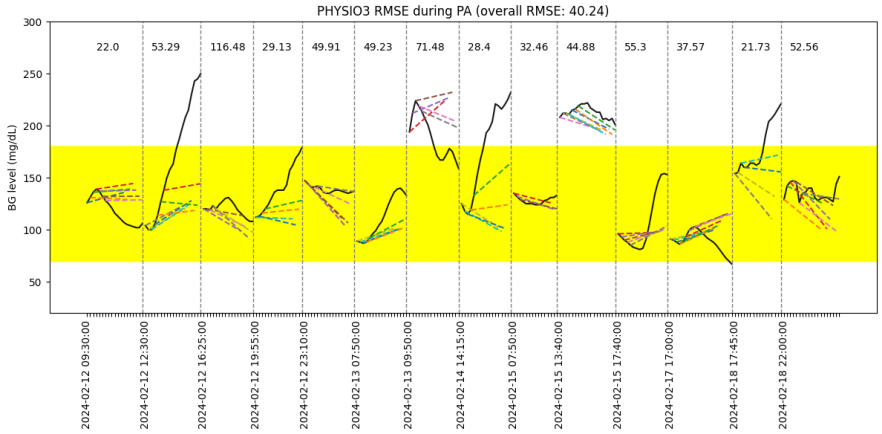

Figure 4.34: Trajectory plot on LSTM-based Physiological Hybrid Model for Participant 2

Despite the overall RMSE being slightly better than that of the benchmark model, Figure 4.33 indicates that the prediction trajectories of the TCN-based Physiological Hybrid model tend to converge toward a median value. Conversely, Figure 4.34 does not show any abnormal prediction patterns, although the prediction performance has degraded compared to the benchmark model.

4.3.2 Ensemble Model Performance During Physical Activity

To see if the ensemble approaches improve the prediction accuracy, three different ensemble models were tested to see the improvement in BG prediction during the PA. All of them were stacked but with different base models. From here on, the models will be addressed

as Stacked Model 1, Stacked Model 2, and Stacked Model 3 for simplicity and clarity. The details of each model are explained in the Methods chapter.

Root Mean Squared Error

Figure 4.35 depicts the mean differences between the RMSE of stacked models' predictions and that of the benchmark model. As explained in the previous section, the benchmark model was chosen based on the dataset for each participant. For Participant 1, the TCN model was selected. For Participant 2, the stacked model combining MLP and PLSR was selected. They are compared with three distinct ensemble models, all of which are stacked.

A negative value in the bar chart indicates that the RMSE of the stacked model is lower than the benchmark model, suggesting better performance. The results reveal that Stacked Model 1 achieved the best performance during physical activity for Participant 2's data, while Stacked Model 2 showed superior results overall, with lower RMSE, MAE, and higher scores on the Clarke Error Grid and Parkes Error Grid. Stacked Model 3 performed best during physical activity for Participant 1's data and also had the best overall RMSE, MSE, and scores on the Clarke Error Grid and Parkes Error Grid.

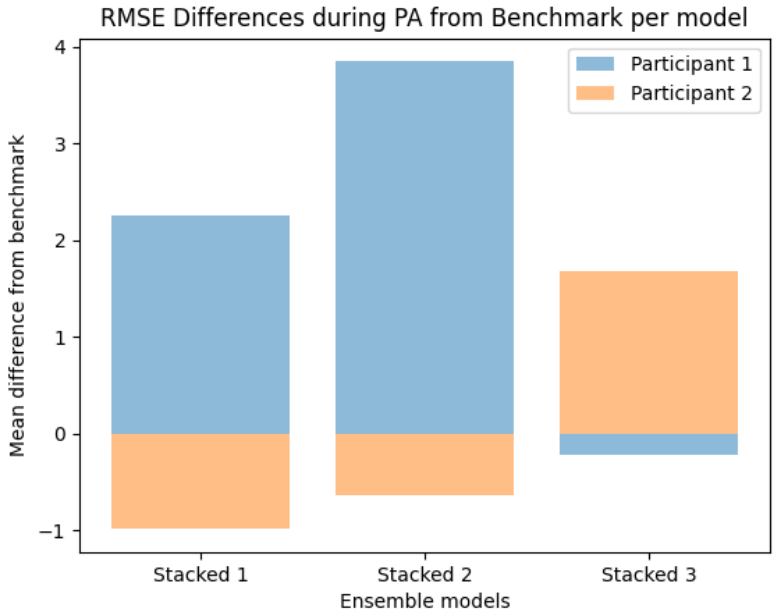

Figure 4.35: Bar chart on Mean Differences from Benchmark Model

Trajectory Plot During PA

Below are the trajectory plots comparing the best-performing stacked model with the benchmark model during physical activity.

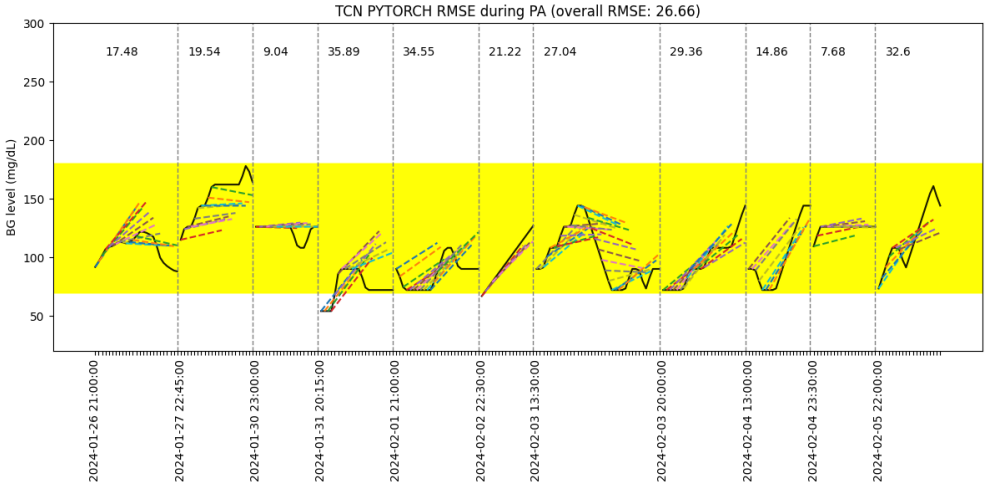

Figure 4.36: Trajectory plot on Benchmark Model (TCN) on Participant 1 data

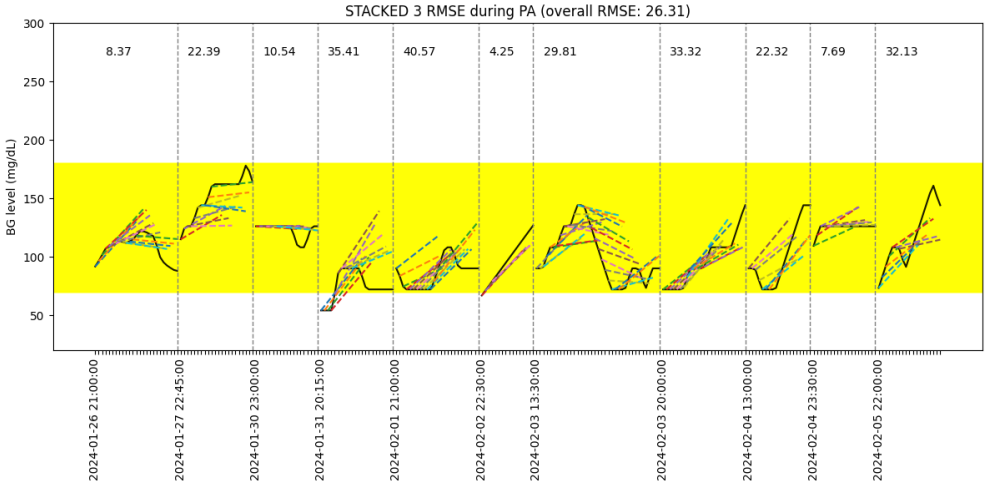

Figure 4.37: Trajectory plot on Stacked Model 3 on Participant 1 data

It should be noted that the average RMSEpa for Stacked Model 3 on Participant 1 data is 22.43 while the average RMSEpa for TCN on Participant 1 data is 22.66. As the mean difference is small, there is no significant variation in the trajectory patterns between Figure 4.36 and Figure 4.37.

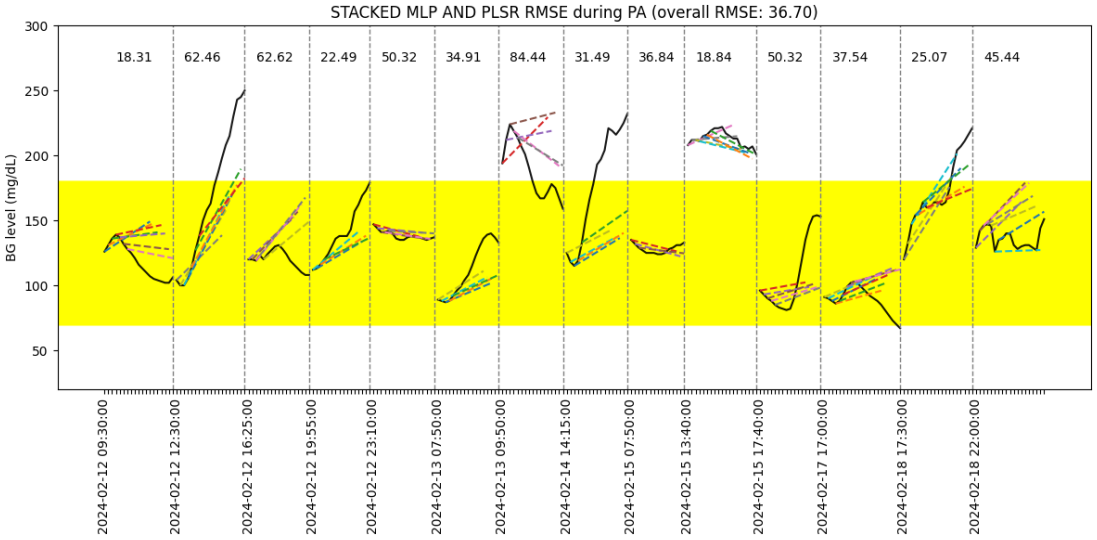

Figure 4.38: Trajectory plot on Benchmark model (Stacked MLP and PLSR) on Participant 2 data

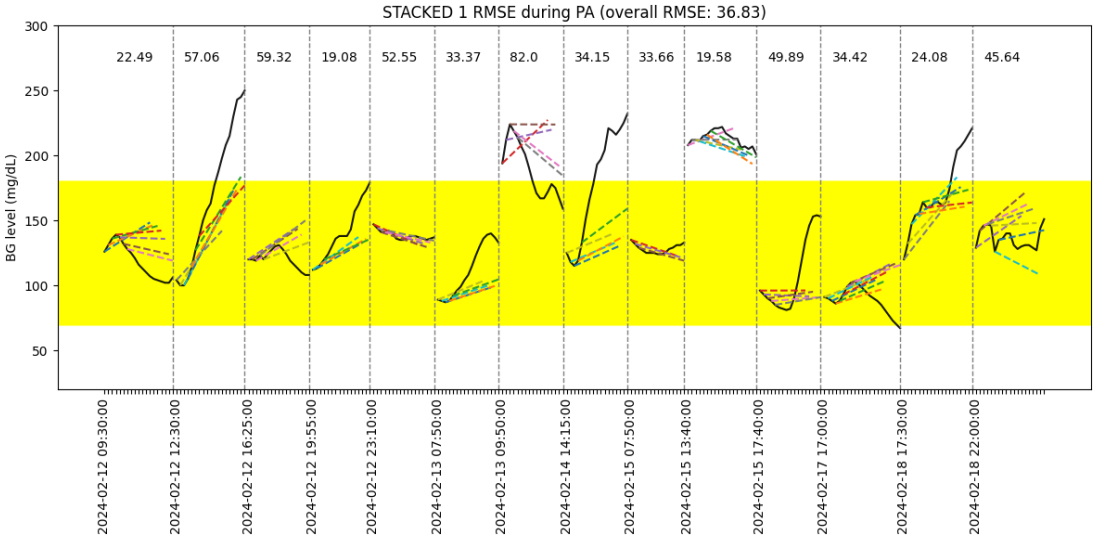

Figure 4.39: Trajectory plot on Stacked Model 1 on Participant 2 data

The average RMSEpa for Stacked Model 1 on Participant 2 data is 40.51, while the benchmark average RMSEps is 41.50. Again due to the slight improvement, there is no significant difference between Figure 4.38 and Figure 4.39

5

Discussion

5.1 Discussion on Comparison Work Result

This chapter mirrors the structure of the Results chapter, discussing each result, including a comparison study, an analysis of predictions made by multiple ML models using data collected from two participants, and the outcomes derived from employing different approaches to enhance predictions during physical activity (PA). Additionally, I have added a discussion that delves into alternative metrics for evaluating blood glucose (BG) levels during PA. Lastly, I end my discussion with the limitations of my study.

5.1.1 Part 1 - Validation with the same dataset

The observed differences between the reported and actual results were analyzed to determine their practical significance. Both Paper 2 and Paper 4 presented reported values that were remarkably close to the actual results obtained during the validation and experimentation process. Paper 2 showed strong agreement with the values reported in the paper across patients. The results obtained from the Paper 4 source code also closely mirror the reported RMSE values for Multi-task learning (MTL) and Single-task learning (STL) at both 30 and 60-minute prediction horizons. The observed RMSE and Mean Absolute Error (MAE) were very close to the reported values, indicating a high degree of consistency and accuracy. The similarity between the reported and actual results validates the robustness and reliability of the models presented in these papers. This agreement indicates the consistency and reproducibility of the models' performance as highlighted in the research papers.

In contrast, papers 1, 3, and 5 showed slight discrepancies between their reported and experimentally derived results. In particular, in paper 3, the reported RMSE values for different experimental setups showed slight deviations from the observed results. The actual RMSE values consistently tended to be slightly higher than the reported values, with an average discrepancy of approximately 1.02 (mg/dL). This variance extended to 1.81(mg/dL) between reported and actual results.

In addition, the RMSE values obtained from Paper 5 showed variability compared to the reported results. Across different feature sets and patient groups, the actual RMSE values tended to be slightly higher than the reported values. These comparisons highlighted inconsistencies between reported and actual results, underscoring the critical need for rigorous validation and verification of research findings. Such discrepancies indicate potential variations in model performance or nuanced experimental conditions that require careful validation practices in research settings.

Nonetheless, all of the models have shown an insignificant amount of difference, validating the reported performance.

5.1.2 Part 2 - Validation with a different dataset

During the validation processing with an Apple Health dataset, among the five papers examined, as shown in the Result chapter, models from paper 1 to paper 3 produced results consistent with each other, while the model from paper 4 exhibited notably inferior performance. This discrepancy may suggest potential inadequacies in implementing the necessary modifications to accommodate the new dataset effectively. It is also interesting to note that the model from the second paper actually generated slightly better performance when applied to the Apple Health dataset. Although there was a challenge in adapting the model proposed in the fifth paper, the validation process highlighted the necessity of comprehensive documentation of hyperparameters and model configurations in the reported papers in order to generate reproducibility and reliability in research outcomes. Due to the difficulty in directly modifying the code for the validation work for the model in the 5th paper, and also for the better scalability of handling different datasets for different models, I have implemented these models into the GluPredKit platform as the next step of my comparison work.

5.1.3 Part 3 - Implementation in GluPredKit

While the outcomes of models from paper 1 to paper 3 showed similar performance values even after implementing the model in the GluPredKit framework, the remaining models from the two papers showed inferior performance, with distinctly poorer results. The implementation of LSTM models (paper 1 to paper 3) posed notable challenges attributable to their

inherent sensitivity to hyperparameters, necessitating fine-tuning procedures. It is noteworthy that, despite replicating the model exactly as delineated in Paper 1, it consistently yielded markedly inferior performance outcomes, indicating the need for preprocessing of the data. Furthermore, the endeavor to replicate the code from paper 5 proved to be exceedingly arduous, primarily owing to its initial development in MATLAB. Moreover, discrepancies emerged from using different libraries, some of which were not fully compatible with their Python equivalents. For example, the ReliefF library in MATLAB may not work seamlessly with its Python counterpart, skrebate. To focus on the primary goal of my thesis—developing a machine learning model that performs well during physical activity—I decided to move on to the next task. It should also be noted that each paper used distinct preprocessing strategies, which likely resulted in significant variations in performance, even when the same model was used. Nonetheless, it would have been more interesting to develop a greater variety of preprocessing methods on the GluPredKit platform.

5.1.4 Comparison Work Conclusion

In conclusion, the results reported in all the papers could be validated by running the open-source code (comparison work part 1). Additional validation was conducted with a different dataset (comparison work part 2), but I acknowledge failing to replicate the results for Paper 4 and Paper 5. The model was subsequently implemented within the GluPredKit framework to ensure scalability for applications across different datasets (comparison work part 3). The variation of the model from Paper 2 produced good results, which were used for the subsequent work on developing a model that performs well during physical activity. The original model in Paper 2 is a stacked model comprising three base regressions: MLP, LSTM, and PSLR. The modified version of this model uses two of these, MLP and PSLR. This altered model was even used as a benchmark for Participant 2's data due to its strong performance in RMSE, MAE, Clarke Error Grid, and Parkes Error Grid.

5.2 Discussion on Working with Real Data

This section presents an analysis of overall model performance, highlighting the best-performing models: the Ridge model and a stacked MLP and PLSR model. This section also includes detailed observations during and after physical activity.

5.2.1 Overall Performance Analysis

The overall performance analysis identified that the Ridge model performed the best on participant 1's data and the stacked MLP and PLSR model performed the best on participant 2's data. Across both datasets, Ridge, ARX, and Stacked MLP and PLSR models have

consistently appeared in the top 5 for the overall performance score. This consistency may suggest the robustness of these algorithms, although they were tested only on data from two participants. However, the strong performance of the conventional linear regression models, Ridge and ARX, indicates that linear regression remains relevant despite the emergence of more complex models.

5.2.2 During Physical Activity Outcome Analysis

What's intriguing from the box plot results is that during the 60-minute prediction horizon (PH), the TCN performed exceptionally well with the mean RMSEpa (RMSE during physical activities) of 22.66, with the best overall RMSE and minimal variabilities for Participant 1's data. Conversely, the LSTM model achieved the best mean RMSEpa of 34.55 for Participant 2's data, but with the second-worst overall RMSE score. This suggests that deep learning models consistently excel in RMSEpa, irrespective of their overall RMSE score. Thus, just by looking at the box plots, it seemed to suggest the potential promise or at least the robustness of deep learning models in accurately capturing physiologically variable situations such as during physical activities.

However, during the analysis of the trajectory plot for the LSTM model on Participant 2 data, abnormal behavior was detected, indicating a consistent pattern of predicting the same blood glucose level value for all predictions. This discovery from the trajectory plots provided a crucial insight that while RMSE is often used to assess machine learning model performance, it may not fully disclose the predictive behavior of the model. This exemplified how solely relying on RMSE for evaluating the model's performance could be misguiding, especially in scenarios involving potential hypoglycemia. Additionally, trajectory plots have unveiled a shortfall of the experimented models some model struggles to grasp the underlying pattern of blood glucose regulations.

Plus, trajectory plots demonstrated the usefulness of visualizing a model's prediction behavior in detail. They can be used to check whether a model has been properly trained, as they reveal anomalies such as always predicting the median value. It is an intuitive tool that allows one to assess a model's performance by examining how closely the predicted trajectory aligns with the actual BG levels. Furthermore, trajectory plots illustrate that most models tend to revert to a safe range. Ideally, a well-trained model should accurately predict even near hyperglycemic and hypoglycemic events.

According to *No Free Lunch Theorem* [95], there is no such a learning algorithm that outperforms any other learning algorithms in every problem set. Despite this, it was both disappointing and intriguing to find that deep learning models, such as LSTM and TCN, did not significantly outperform classic regression models in predicting blood glucose levels for people with T1DM. However, it should be noted that only three months of real data were used to train each model for the comparison. This limited data may not have been sufficient

to account for the complexity of the model.

Nonetheless, a benchmarking study conducted by Xie et al. compared machine learning algorithms for blood glucose prediction with classical time-series models. The study concluded that no model consistently outperformed the classical ARX model. However, TCN exhibited robustness in tracking blood glucose trajectories with sporadic oscillations, whereas the ARX model tended to over-predict peak blood glucose levels and under-predict valley levels [51]. This study aligns with my findings, which indicate the absence of a clear winner among the models. This consistency resonates with the broader understanding that no single model emerges as the single best performer in predicting blood glucose levels. Nonetheless, the study still highlights that TCN demonstrated robustness in tracking BG levels with sporadic oscillations. This suggests that deep learning models remain relevant and hold the potential for addressing the complexities of blood glucose prediction in individuals with T1DM.

In addition to highlighting the potential of deep learning models in addressing the complexities of blood glucose prediction for individuals with T1DM, it's crucial to acknowledge the challenges associated with these models. For instance, A. Casolaro et al [72] point out limitations associated with deep learning architectures for time series forecasting. Besides the confidence interval estimation issue, it points out that when the deep learning architectures become increasingly complex, they become more susceptible to overfitting which undermines the robustness of the models and compromises their ability to generalize well to new data. Plus it points out that some deep learning architectures, such as Transformers, require a large number of parameters to be estimated, necessitating adequate long time series data for training. While data augmentation techniques partially mitigate this challenge, current solutions are not fully satisfactory. These challenges underscore the importance of further research and development efforts to enhance the performance and applicability of deep learning models in the context of blood glucose prediction for individuals with T1DM.

5.2.3 After Physical Activity Outcome Analysis

No clear patterns emerged from the bar chart and trajectory plots. This could be because the body's physiological processes are very complex, and there are many variables besides physical activity that affect blood glucose levels.

While it's undeniable that physical activity affects blood glucose levels, factors like the intensity and duration of the activity may contribute to the lack of consistent patterns. This reinforces the idea that predicting blood glucose levels solely based on physical activity is difficult due to the complexity of these processes.

In summary, the interaction of various physiological processes and other factors complicates the identification of meaningful patterns in blood glucose levels immediately after physical

activity. However, it may be necessary to adopt a more comprehensive or creative approach to uncover patterns within the data regarding how the models have predicted after physical activities.

5.3 Discussion on Result from Approaches Taken to Improve Predictions during Physical Activities

This section discusses the result from the application of two approaches to improve blood glucose (BG) level predictions during physical activity: the Physiological Hybrid Model and various ensemble models.

5.3.1 Application of Physiological Hybrid Model

The application of the Physiological Hybrid model resulted in poorer performance compared to the benchmark models. Particularly disappointing was the outcome of the TCN-based Physiological Hybrid Model on Participant 1 data, as the overall RMSE actually slightly improved compared to that of the benchmark model. What was evident from the trajectory plot of the TCN-based Physiological Hybrid model was a tendency for the model to predict toward the median value most of the time, suggesting that the model may have failed to accurately capture physiological patterns to predict blood glucose levels. This may indicate a need for more data to effectively capture the complex physiological relationships between the inputs, implying that the current dataset is insufficient to mitigate the problem of overfitting. The LSTM-based Physiological Hybrid Model performed poorly on Participant 2 data, exhibiting shortcomings in both overall RMSE and RMSE during physical activity. Further investigation is necessary to gain a better understanding of the physiological hybrid approaches.

5.3.2 Application of Ensemble Model

The use of ensemble models was initially aimed at exploring whether they could enhance BG level predictions during PA. At the outset, there was an expectation that combining multiple models would lead to improved performance, leveraging the strengths of each individual model. They did improve performance. However, the findings revealed that no single ensemble model consistently outperformed the others, and the optimal model varied depending on the dataset. Stacked Model 1 achieved the best results for data from Participant 2, while Stacked Model 3 was most effective for data from Participant 1.

It's intriguing that Stacked Model 1 achieved the best results for data from Participant 2,

even though this participant's data had relatively high oscillation in BG. This outcome was unexpected, especially considering that Stacked Model 1 did not include any deep learning components, which are typically assumed to excel in identifying complex patterns. By contrast, the other two ensemble models did have deep learning components as a base model, with Stacked Model 2 incorporating an MLP and Stacked Model 3 utilizing a TCN. This observation led to speculation that deep learning models might struggle with generalization, particularly in the presence of noisy data. It raises the intriguing possibility that simpler models, with their reduced complexity, may offer advantages in certain scenarios by avoiding overfitting issues.

Reflecting on these findings, it becomes apparent that the performance of ensemble models is not solely determined by the individual components but also by their interactions and adaptability to the specific characteristics of the data. While uncertainties remain regarding the optimal combination of base models, the overall improvement observed in BG predictions during PA reaffirms the relevance and potential of ensemble modeling in addressing the complexities of BG prediction in individuals with T1DM.

5.4 Other Methodologies for Evaluating Performance BG during PA

In this section, I have explored various metrics and methodologies for evaluating the performance of BG predictions during PA. While RMSE is commonly used, its limitations necessitate alternative approaches that could provide a more comprehensive understanding of model behavior and prediction accuracy. To offset, I did use trajectory plots in the project but there are diverse approaches that can lead to a more comprehensive assessment of prediction performance. For instance, the J index can indicate the consistency and clinical usefulness of a prediction. The J index is calculated using the Error Sum of the Squared Differences (ESOD) and the Time Gain (TG). The Mean Amplitude of Glycemic Excursions (MAGE) evaluates glycemic variability. The simulation capabilities of the ReplayBG methodology allow for exploring various therapy scenarios, adding a dynamic aspect to the evaluation process.

5.4.1 The Limitations of Root Mean Squared Error in Capturing Model Prediction Behavior

The dynamics of blood glucose levels during physical activities are notably intricate, influenced by factors such as insulin sensitivity, carbohydrate intake, stress levels, and individual metabolic variations. These multifaceted dynamics often elude traditional machine learning models resulting in disparities between predicted and actual glucose trajectories, even with

the models that have resulted with lower RMSE values.

5.4.2 Other Evaluation Approaches

In the realm of continuous glucose monitoring (CGM) glucose prediction algorithms, the quest for an optimal parameter set and effective comparison of different prediction strategies poses significant challenges. Facchinetti et al [90] argue that a new index is required to compare continuous glucose monitoring (CGM) glucose prediction algorithms because existing criteria and methodologies are not sufficient for solving the problems of finding the optimal parameter set and comparing different predictions strategies effectively.

Recognizing these limitations, the need for a comprehensive evaluation index becomes apparent. Such an index should simultaneously consider the regularity of predicted profiles and the time gained due to prediction. Addressing this need, the development of a new index, denoted as J, is proposed. This index combines the Error Sum of the Squared Differences (ESOD) to measure regularity and Time Gain (TG) to quantify the time gained through prediction. The paper [90] has claimed that by normalizing ESOD and TG within the J index, a balanced evaluation of prediction performance is achieved, facilitating the optimal design and comparison of glucose prediction algorithms and there are papers [91] [92] which have used these evaluation metrics to measure the performance.

I think using ESOD and TG can provide insights on how reliable the predictions are as they measure regularity and time gain which makes it more intuitive how useful the predictions for patients with diabetes are.

ESOD (Error Sum of the Squared Differences) ESOD serves as a measure of the regularity of predicted glucose profiles. High ESOD values indicate irregularities and spurious oscillations in the predicted profile, potentially leading to false alerts and reduced clinical utility. Monitoring ESOD aids in assessing the reliability and stability of predicted glucose profiles, ensuring smooth predictions free from unnecessary fluctuations. The ESOD formula from [90] is:

$$ESOD(x) = \frac{1}{N}\sum_{i=1}^{N}(\Delta^2(x(i)))^2$$

The normalized ESOD formula used in [92] is as follow:

$$ESODn = \frac{\sum_{k=3}(\hat{y}(k) - 2\hat{y}(k-1) + \hat{y}(k-2))^2}{\sum_{k=3}(y(k) - 2y(k-1) + y(k-2))^2}$$

[92] states that the best ESODn value is one. The closer it gets to one, the better the predicted time series are considered. I think the value of using ESODn is that the value of ESODn shows how the predictions fluctuate unnecessarily compared to the actual blood glucose

levels. This metric identifies models that might trigger false alarms or miss significant changes in blood sugar levels. By assessing these unnecessary fluctuations, ESODn helps ensure that the prediction model is more stable and less likely to generate false alerts for low or high blood sugar events. This is especially useful because hypoglycemia is more likely to occur during PA.

TG (Time Gain) TG evaluates the time gained through prediction by comparing predicted glucose profiles with actual data. It is a valuable metric for evaluating a blood glucose level prediction model's performance, especially during physical activity, because it measures how much lead time the prediction provides before hypo- or hyperglycemic events occur. High TG values suggest that the algorithm can give early warnings, enabling timely interventions to prevent extreme fluctuations in blood glucose levels. In contrast, low TG values indicate that the predictions are too close to the actual event, limiting the opportunity for proactive measures. Monitoring TG, therefore, offers a more practical insight into the prediction model's effectiveness in providing actionable alerts, which is crucial for managing blood glucose during physical activity, where fluctuations can occur rapidly. This approach offers a more clinical perspective than RMSE, which only measures average error without considering the timing of predictions.

Corresponding to a sampling time Δt (minute), a total of N samples, and an L-step ahead prediction horizon, TG (minute) is defined as follows:

$$\text{delay} = \arg\min_{k\in[0,\text{PH}]} \left\{ \frac{1}{N-\text{PH}} \sum_{k=1}^{N-\text{PH}} \left(\hat{y}(k+i) - y(k)\right)^2 \right\}$$

$$TG = (L - \text{delay}) \cdot \Delta t$$

Where $\hat{y}(k+i)$ denotes the L-step ahead prediction based on $y(k+i-L)$, and the "delay" measures the temporal shift that minimizes the distance between the prediction and the actual value of the blood glucose level. A larger TG implies an earlier detection of a potential hypo/hyperglycemia event, while zero-order-hold prediction will render $TG = 0$ and thus is not useful from a clinical perspective [92].

J Index The J index combines ESOD and TG to offer a comprehensive evaluation of predicted profiles. By normalizing ESOD to the irregularity of the original data and TG to the prediction horizon, the J index provides a balanced assessment of regularity and time gain achieved through prediction. A lower J value indicates a more consistent and clinically useful prediction, making it a valuable tool for optimizing parameter sets and comparing different prediction methods. It aligns with the interpretation that a good prediction should have a low ESOD (to be reliable) and a high TG (to be clinically useful) [90]. The J index can serve as a reliable criterion for selecting the optimal parameter set and designing effective

glucose prediction algorithms.

$$J = \frac{\frac{ESOD(\hat{x})}{ESOD(x)}}{\frac{TG(x,\hat{x})}{PH}}$$

Use of Mean Amplitude of Glycemic Excursions (MAGE) The Mean Amplitude of Glycemic Excursions (MAGE) is a crucial index utilized to evaluate glycemic variability, particularly in individuals with diabetes. It is derived as the arithmetic average of upward or downward excursions in blood glucose levels that surpass a predefined threshold, usually determined by the standard deviation of blood glucose concentrations over a 24-hour period.

It is computed based on the optimal sequence of extreme points. Two cases are considered, the case when the first extreme point is a local minimum and the case when it is a local maximum. Then MAGE is calculated as the sum of positive and negative excursions around the extreme points, and an average is taken.

MAGE could be also used to train a model on the variabilities of blood glucose levels. Syafaah et al. [93] utilized MAGE as a metric to assess glycemic variability in individuals. These MAGE values were then used as features in a machine-learning model to predict the diabetes status of the patients.

While RMSE measures the average error between predicted and actual blood glucose levels, it doesn't differentiate between consistent errors and wide-ranging fluctuations. In contrast, MAGE focuses on the amplitude of these excursions, offering a better understanding of how much and how frequently blood glucose levels deviate from the norm. MAGE is a more clinically relevant metric, especially during physical activity when blood glucose levels can fluctuate rapidly. Thus, I think using MAGE as an evaluation metric can offer a more precise understanding of the risks linked to glycemic variability.

A Digital Twin-Based Methodology ReplayBG [94] introduced a digital twin-based methodology to identify a personalized model from type 1 diabetes data and simulate glucose concentrations to assess alternative therapies. The ReplayBG methodology is used to evaluate the effectiveness of different insulin and carbohydrate therapies for simulating glucose concentrations in individuals with type 1 diabetes (T1D).

The methodology involves two main steps: first, a personalized model of glucose-insulin dynamics is identified using data on insulin, carbohydrate intake, and continuous glucose monitoring. Second, this model is used to simulate glucose concentrations under different therapy scenarios. The study evaluated ReplayBG on virtual subjects and compared its performance with existing methods, showing high accuracy in simulating the effects of treatment alterations.

I think this could be further developed to include factors such as PA to enhance the person-

alized modeling of glucose-insulin dynamics in individuals with T1DM. By incorporating a subsystem that describes physical exercise, the model could better capture the impact of varying activity levels on glucose dynamics. This expansion would allow for a more comprehensive assessment of how different lifestyle factors, including PA.

5.5 Limitations

The following are the limitations of my work. The major constraints include a small number of participants and the inherent challenges of collecting accurate data with smartwatches. Additional limitations relate to the nature of the machine learning modeling approaches used.

5.5.1 Use of Smartwatches to Collect Data

I acknowledge that there are inherent limitations of smartwatches, including potential inaccuracies in data measurements and individual variations in user responses.

Smartwatches, while providing a convenient means of monitoring physical activity and health parameters, may exhibit discrepancies in accuracy, particularly concerning metrics such as heart rate monitoring and step counting. Variability in sensor performance, placement on the wrist, and user activity levels can contribute to inconsistent data readings, leading to potential inaccuracies in the collected data. This study found that smartwatch-based heart rate variability (HRV) measurements are less accurate than traditional HRV methods, such as electrocardiogram (ECG). The Błaszczykowski et al. [65] attribute this inaccuracy to factors such as sensor placement and movement artifact. O'Connor et al. [66] suggest that smartwatches may be less accurate for measuring steps in certain populations, such as older adults or individuals with obesity. Nonetheless, Chen et al. [67] found that smartwatch-based heart rate monitoring accuracy is generally good during exercise while acknowledging that it can be still affected by factors such as sensor placement, device type, and activity intensity.

These limitations have implications for the interpretation of results obtained from smartwatch-derived data. Inaccurate or unreliable measurements may introduce bias and affect the validity of conclusions drawn from the dataset. Moreover, individual variations in user responses, such as differences in physiological characteristics, fitness levels, and adherence to wearing the device, further compound the challenge of data interpretation [68].

Although resolving the limitations of the measurement tools is beyond the scope of this thesis, imputation was used during data preprocessing to address quality issues resulting from smartwatch usage, such as missing data due to low battery or device removal. It's

important to note that data from both participants were collected using smartwatches: one used a Fitbit for step counts, while the other used an Apple Watch for various metrics including heart rate, step counts, and calories burned.

5.5.2 Limited Amount of Participants

For this work, data from only two participants were used for reasons of convenience and availability. Recruiting participants to collect health data involves considerable time and effort. Including the supervisor's data helped save a significant amount of both. It's also worth noting that the data collection process differed for each of the two participants, which was intended to ensure the authenticity of data acquisition. However, due to the limited amount of participants which still then used different datasets, generalizing the results remains challenging.

5.5.3 Ensemble Model Approach

Although a slight improvement in prediction performance was observed by applying ensemble models, there are many possible combinations of base models, making it uncertain which approach is the best. This continues to be a challenge for researchers seeking to find the optimal combinations of techniques [82].

5.5.4 Deep Learning Models

Deep learning architectures can be prone to overfitting and require substantial data for training due to their complexity. The underperformance of deep learning-based physiological hybrid models in this study might be attributed to their vulnerability to generalization issues, especially when working with noisy data. Simpler models might be more robust under such conditions. Additionally, identifying the optimal hyperparameters for deep learning models proved to be challenging.

6 Conclusion

Based on the literature review conducted during the early stage of my thesis, I have aimed to identify methods that enhance model performance and accuracy during physical activity. To enhance the BG level prediction during the PA for people with T1DM, data collected from two participants with T1DM was used to train diverse models. These are evaluated in various ways. Below is a summary of the key findings from the process, and suggestions for future research to achieve better outcomes.

6.1 Summary of Key Findings

The GluPredKit platform proved valuable in implementing models and scaling across various datasets. The platform offered a more standardized approach to data preprocessing, model implementation, and evaluation process which contributed to the consistency of results.

Also, one of the main discoveries of this study is that using only RMSE to assess a model's prediction performance might not give a full understanding, or even misguiding in some cases. Trajectory plots were particularly effective in preventing such issues, showing prediction patterns and illustrating how models perform over time. Additionally, other evaluation metrics such as ESOD, TG, J index, Clarke Error Grid, and Parkes Error Grid could be utilized. However, employing a comprehensive approach is crucial for better assessing the quality of the model's predictions.

Another finding is that no clear pattern emerged regarding how prediction performance changed during the shifted period (the interval from a certain number of minutes after the physical activity to the end of the activity plus additional skewed minutes) or after a certain interval of the physical activity. The underlying physiological processes might be too complex to exhibit any consistent trends.

Although deep learning models such as LSTM and TCN are extensively utilized, they did not demonstrate a significant advantage over traditional regression models in forecasting blood glucose levels for individuals with T1DM. This could be attributed to either insufficient data input or inadequate hyperparameter tuning that could potentially enhance the models. Nevertheless, it is noteworthy that these complex models do not surpass conventional ML regression models which then proves that they continue to hold relevance.

Lastly, to summarise the findings from implementing approaches to improve predictions during PA, the Physiological Hybrid model performed worse than the benchmark models during physical activity, indicating the need for additional data or architectural changes to address the complex physiological relationships affecting predictions. On the other hand, ensemble models showed improved performance during PA but with varying results, with no single model consistently outperforming the others. Stacked Model 1 yielded the best results for data from Participant 2, while Stacked Model 3 proved the most effective for Participant 1, indicating that model performance can depend heavily on the dataset and context.

6.2 Research Contribution

Throughout the research, it was demonstrated that the development and implementation of the GluPredKit platform provided scalable solutions across various models and datasets. This capability allows for easier comparisons between multiple models and across different datasets, potentially improving consistency and reproducibility in future studies on blood glucose machine-learning predictions.

The key findings from this work revealed that evaluating a model solely based on RMSE has limitations. They highlighted the need for a more comprehensive evaluation approach to guide further research in improving blood glucose predictions during and after physical activity.

Despite their popularity, deep learning models like LSTM and TCN did not significantly outperform traditional regression models in predicting blood glucose levels for people with T1DM. This finding emphasizes the importance of not over-relying on complex models when simpler ones might achieve similar results, suggesting that researchers should critically assess the need for deep learning in specific scenarios.

Although ensemble models showed modest performance improvements during physical activity, their results varied, suggesting that no single ensemble model consistently outperforms the others. This finding highlights the importance of carefully choosing base models and the impact that input data can have on the final outcomes.

6.3 Future Work

The future direction of this study could focus on expanding the participant pool and exploring advanced evaluation metrics to improve the generalizability and reliability of the blood glucose prediction models during physical activity (PA). Ultimately, the goal is to create a model that more accurately captures the physiological dynamics during PA.

6.3.1 Recruiting More Participants and Expanding the Scope

When the scope of participants is expanded, the results of this study could become more generalizable, allowing for broader applications of the findings. A larger and more diverse participant pool can help reveal more meaningful patterns, offering insights into various factors that may influence the results. This expansion can lead to increased reliability by reducing biases that may be in the current work, and the increased reliability will then draw stronger conclusions.

6.3.2 Finding an optimal measurement to assess the performance of a model during PA

Although it was theoretically discussed about other useful measurement methods in the Discussion chapter, identifying meaningful metrics to evaluate the machine learning models, particularly during PA, still remains an area for future work. In this study, the RMSE during PA was calculated, along with analyses of trajectory plots. However, metrics such as the J index and MAGE could further enhance the evaluation, as they account for the variability in blood glucose levels, which are often subject to significant fluctuations during physical activity.

6.3.3 Achieving a model that understands physiological dynamics during PA

Using an ensemble model improved overall prediction performance, including predictions during PA. However, to truly better capture the underlying physiological dynamics during PA, I think it is worth exploring physiological hybrid approaches. The premise is that a more

profound understanding of physiological processes should lead to more accurate predictions of blood glucose levels during physical activity.

6.4 Closing

Despite the challenges and variability in model performance, I believe that I have achieved a broader understanding of blood glucose prediction during physical activity and highlighted the ongoing need for rigorous validation and exploration of innovative approaches to improve blood glucose prediction during physical activities for people with T1DM.

Bibliography

[1] C. Colom, A. Rull, J. L. Sanchez-Quesada, and A. Pérez, “Cardiovascular Disease in Type 1 Diabetes Mellitus: Epidemiology and Management of Cardiovascular Risk,” Journal of Clinical Medicine, vol. 10, no. 8, p. 1798, Apr. 2021, doi: https://doi.org/10.3390/jcm10081798.

[2] S. G. Albert and M. Bernbaum, “Exercise for Patients With Diabetic Retinopathy,” Diabetes Care, vol. 18, no. 1, pp. 130–132, Jan. 1995, doi: https://doi.org/10.2337/diacare.18.1.130.

[3] Bird SR, Hawley JAUpdate on the effects of physical activity on insulin sensitivity in humansBMJ Open Sport & Exercise Medicine 2017;2:e000143. doi: 10.1136/bmjsem-2016-000143

[4] S. J. Livingstone et al., “Risk of Cardiovascular Disease and Total Mortality in Adults with Type 1 Diabetes: Scottish Registry Linkage Study,” PLoS Medicine, vol. 9, no. 10. Public Library of Science (PLoS), p. e1001321, Oct. 02, 2012. doi: 10.1371/journal.pmed.1001321.

[5] R. Sealand, C. Razavi, and R. A. Adler, “Diabetes Mellitus and Osteoporosis,” Current Diabetes Reports, vol. 13, no. 3. Springer Science and Business Media LLC, pp. 411–418, Mar. 08, 2013. doi: 10.1007/s11892-013-0376-x.

[6] E. J. Hamilton et al., “Prevalence and predictors of osteopenia and osteoporosis in adults with Type 1 diabetes,” Diabetic Medicine, vol. 26, no. 1. Wiley, pp. 45–52, Jan. 2009. doi: 10.1111/j.1464-5491.2008.02608.x.

[7] Borg WP, Sherwin RS, During MJ, Borg MA, Shulman GI. Local ventromedial hypothalamus glucopenia triggers counterregulatory hormone release. Diabetes. 1995;44:180–184.

[8] G. E. Umpierrez and A. E. Kitabchi, “Diabetic Ketoacidosis,” Treatments in Endocrinology, vol. 2, no. 2, pp. 95–108, 2003, doi: https://doi.org/10.2165/00024677-200302020-00003.

[9] C Marling and R. Burnescu, 'The ohiot1dm dataset for blood glucose level prediction: Update 2020', 2019. http://smarthealth.cs.ohio.edu/bglp/OhioT1DM-dataset-paper.pdf (accessed Aug. 31, 2023).

[10] A. Zaitcev, M. R. Eissa, Z. Hui, T. Good, J. Elliott, and M. Benaissa, "Automatic inference of hypoglycemia causes in type 1 diabetes: a feasibility study," Frontiers in Clinical Diabetes and Healthcare, vol. 4. Frontiers Media SA, Apr. 17, 2023. doi: 10.3389/fcdhc.2023.1095859.

[11] Afentakis, I., Unsworth, R., Herrero, P., Oliver, N., Reddy, M., & Georgiou, P. (2023). Development and Validation of Binary Classifiers to Predict Nocturnal Hypoglycemia in Adults With Type 1 Diabetes. J Diabetes Sci Technol, p. 19322968231185796. doi: 10.1177/19322968231185796

[12] Balakrishnan, N. P., Rangaiah, G. P., & Samavedham, L. (2012). Personalized blood glucose models for exercise, meal and insulin interventions in type 1 diabetic children. In 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 1250–1253). doi: 10.1109/EMBC.2012.6346164

[13] Bergford, S., et al. (2023). The Type 1 Diabetes and EXercise Initiative: Predicting Hypoglycemia Risk During Exercise for Participants with Type 1 Diabetes Using Repeated Measures Random Forest. Diabetes Technol Ther, Jun. 2023. doi: 10.1089/dia.2023.0140

[14] 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." Accessed: Feb. 28, 2024. [Online]. Available: https://ceur-ws.org/Vol-2148/paper14.pdf

[15] Bertachi, A., et al. (2020). Prediction of Nocturnal Hypoglycemia in Adults with Type 1 Diabetes under Multiple Daily Injections Using Continuous Glucose Monitoring and Physical Activity Monitor. Sensors, 20(6). doi: 10.3390/s20061705

[16] B. Bogue-Jimenez, X. Huang, D. Powell, and A. Doblas, "Selection of Noninvasive Features in Wrist-Based Wearable Sensors to Predict Blood Glucose Concentrations Using Machine Learning Algorithms," Sensors, vol. 22, no. 9, p. 3534, May 2022, doi: https://doi.org/10.3390/s22093534.

[17] Calhoun, P., Levine, R. A., Fan, J. (2020). Repeated measures random forests (RMRF): Identifying factors associated with nocturnal hypoglycemia. Biometrics, 77(1), 343–351. doi: 10.1111/biom.13284

[18] de Canete, J. F., et al. (2012). Artificial neural networks for closed loop control of in silico and ad hoc type 1 diabetes. Computer Methods and Programs in Biomedicine, 106(1), 55–66. doi: https://doi.org/10.1016/j.cmpb.2011.11.006

[19] Cappon, et al. (2020). A Personalized and Interpretable Deep Learning Based Approach to Predict Blood Glucose Concentration in Type 1 Diabetes. KDH 2020 Knowledge Discovery in Healthcare Data 2020, pp. 75-79

[20] Cescon, M., Renard, E. (2011). Adaptive subspace-based prediction of T1DM glycemia. In 2011 50th IEEE Conference on Decision and Control and European Control Conference (pp. 5164–5169). doi: 10.1109/CDC.2011.6161154

[21] I. Contreras, A. Bertachi, L. Biagi, S. Oviedo, and J. Vehí, “Using Grammatical Evolution to Generate Short-Term Blood Glucose Prediction models.” Accessed: Feb. 28, 2024. [Online]. Available: https://ceur-ws.org/Vol-2148/paper15.pdf

[22] Daniels, J., Herrero, P., Georgiou, P. (2020). Personalised Glucose Prediction via Deep Multitask Networks. J Healthc Inform Res, 4(1), 71–90. doi: 10.1007/s41666-019-00063-2

[23] De Paoli, B., et al. (2021). Blood Glucose Level Forecasting on Type-1-Diabetes Subjects during Physical Activity: A Comparative Analysis of Different Learning Techniques. Bioengineering (Basel), 8(6). doi: 10.3390/bioengineering8060072

[24] S. M. Ewings, S. K. Sahu, J. J. Valletta, C. D. Byrne, and A. J. Chipperfield, “A Bayesian network for modelling blood glucose concentration and exercise in type 1 diabetes,” Statistical Methods in Medical Research, vol. 24, no. 3, pp. 342–372, Feb. 2014, doi: https://doi.org/10.1177/0962280214520732.

[25] Faccioli, S., et al. (2018). Black-box Model Identification of Physical Activity in Type-l Diabetes Patients. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 3910–3913). doi: 10.1109/EMBC.2018.8513378

[26] E. I. Georga, V. C. Protopappas, D. Polyzos and D. I. Fotiadis, "A predictive model of subcutaneous glucose concentration in type 1 diabetes based on Random Forests," 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 2012, pp. 2889-2892, doi: 10.1109/EMBC.2012.6346567.

[27] Georga, E. I., et al. (2012). Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression. IEEE Journal of Biomedical and Health Informatics, 17(4), 734–741. doi: 10.1109/JBHI.2012.2189064

[28] Georga, E. I., et al. (2015). Online prediction of glucose concentration in type 1 diabetes using extreme learning machines. Annu Int Conf IEEE Eng Med Biol Soc, 2015, 3262–3265. doi: 10.1109/EMBC.2015.7319088

[29] E. I. Georga, V. C. Protopappas, D. Polyzos and D. I. Fotiadis, "Online prediction of glucose concentration in type 1 diabetes using extreme learning machines," 2015 37th

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 2015, pp. 3262-3265, doi: 10.1109/EMBC.2015.7319088.

[30] Georga, E. I., et al. (2019). Short-term prediction of glucose in type 1 diabetes using kernel adaptive filters. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4523–4526). doi: 10.1109/EMBC.2019.8857232

[31] Jaloli M, Lipscomb W, Cescon M. Incorporating the Effect of Behavioral States in Multi-Step Ahead Deep Learning Based Multivariate Predictors for Blood Glucose Forecasting in Type 1 Diabetes. BioMedInformatics. 2022; 2(4):715-726. https://doi.org/10.3390/biomedinformatics2040048

[32] Jeon, J., & Lee, J. (2019). Predicting Glycaemia in Type 1 Diabetes Patients: Experiments in Feature Engineering and Data Imputation. IEEE Access, 7, 174124–174131. doi: 10.1109/ACCESS.2019.2957507

[33] Khadem, M. Z., Azad, R., Rabbani, H., & Kiani, A. K. (2023). Blood Glucose Level Time Series Forecasting: Nested Deep Ensemble Learning Lag Fusion. Sensors, 23(1). doi: 10.3390/s23010072

[34] Liu, J., Li, J., & Wu, H. (2018). Enhancing Blood Glucose Prediction with Meal Absorption and Physical Exercise Information. Sensors, 18(2), 481. doi: 10.3390/s18020481

[35] Martínez-Delgado, M. I., et al. (2023). Using Absorption Models for Insulin and Carbohydrates and Deep Leaning to Improve Glucose Level Predictions. Sensors, 23(3). doi: 10.3390/s23030743

[36] S. Mirshekarian, H. Shen, R. Bunescu and C. Marling, "LSTMs and Neural Attention Models for Blood Glucose Prediction: Comparative Experiments on Real and Synthetic Data," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 706-712, doi: 10.1109/EMBC.2019.8856940.

[37] Mirshekarian, S., & Farrahi, M. (2017). Using LSTMs to learn physiological models of blood glucose behavior. In Proceedings of the 2nd International Conference on Multimedia and Image Processing - ICMAI '17 (pp. 35–39). doi: 10.1145/3060972.3060982

[38] Mosquera-Lopez C, Ramsey KL, Roquemen-Echeverri V, Jacobs PG. Modeling risk of hypoglycemia during and following physical activity in people with type 1 diabetes using explainable mixed-effects machine learning. Comput Biol Med. 2023 Mar;155:106670. doi: 10.1016/j.compbiomed.2023.106670. Epub 2023 Feb 11. PMID: 36803791.

[39] Nemat, H., & Del Prato, S. (2020). Data Fusion of Activity and CGM for Predicting Blood Glucose Levels. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 3230–3233). doi: 10.1109/EMBC44109.2020.9175253

[40] Parcerisas, A., Hernando, M. E., & Barceló, I. (2022). A Machine Learning Approach to Minimize Nocturnal Hypoglycemic Events in Type 1 Diabetic Patients under Multiple Doses of Insulin. IEEE Access, 10, 151881–151891. doi: 10.1109/ACCESS.2022.3144166

[41] J. Pavan et al., "Personalized Machine Learning Algorithm based on Shallow Network and Error Imputation Module for an Improved Blood Glucose Prediction." Accessed: Feb. 28, 2024. [Online]. Available: https://ceur-ws.org/Vol-2675/paper16.pdf

[42] Reddy, M., Herrero, P., & Georgiou, P. (2019). Prediction of Hypoglycemia During Aerobic Exercise in Adults With Type 1 Diabetes. J Diabetes Sci Technol, 13(1), 42–52. doi: 10.1177/1932296818783182

[43] Romero-Ugalde, H. M., & Caicedo-Bravo, E. L. (2019). ARX model for interstitial glucose prediction during and after physical activities. In 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 1–4). doi: 10.1109/BHI.2019.8834545

[44] Sevil, U. G., Cinar, A., & Kahyaoglu, M. (2021). Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management. IEEE J Biomed Health Inform, 25(8), 2928–2937. doi: 10.1109/JBHI.2021.3075605

[45] Shilo, S., et al. (2021). Prediction of Personal Glycemic Responses to Food for Individuals With Type 1 Diabetes Through Integration of Clinical and Microbial Data. Diabetes Care, 44(11), 2699–2706. doi: 10.2337/dc21-0045

[46] Sun, W., Gong, P., & Huang, L. (2021). Prior informed regularization of recursively updated latent-variables-based models with missing observations. Statistica Sinica, 31, 1307–1330. doi: 10.5705/ss.202019.0291

[47] 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, Mar. 2022, doi: https://doi.org/10.1016/j.isci.2022.103888.

[48] Mohammad Reza Vahedi et al., "Predicting Glucose Levels in Patients with Type1 Diabetes Based on Physiological and Activity Data," Jun. 2018, doi: https://doi.org/10.1145/3220127.3220133.

[49] W. P. T. M. van Doorn et al., “Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study,” PLoS ONE, vol. 16, no. 6, p. e0253125, Jun. 2021, doi: https://doi.org/10.1371/journal.pone.0253125.

[50] 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, Jun. 2019, doi: https://doi.org/10.1177/1460458219850682.

[51] J. Xie and Q. Wang, "Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison With Classical Time-Series Models," in IEEE Transactions on Biomedical Engineering, vol. 67, no. 11, pp. 3101-3124, Nov. 2020, doi: 10.1109/TBME.2020.2975959.

[52] Konstantia Zarkogianni et al., “Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring,” vol. 53, no. 12, pp. 1333–1343, Jun. 2015, doi: https://doi.org/10.1007/s11517-015-1320-9.

[53] K. Zarkogianni, K. Mitsis, M. . -T. Arredondo, G. Fico, A. Fioravanti and K. S. Nikita, "Neuro-fuzzy based glucose prediction model for patients with Type 1 diabetes mellitus," IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, Spain, 2014, pp. 252-255, doi: 10.1109/BHI.2014.6864351.

[54] 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, Aug. 2021, doi: https://doi.org/10.1016/j.bspc.2021.102923.

[55] S. N. Scott et al., “Fasted High-Intensity Interval and Moderate-Intensity Exercise Do Not Lead to Detrimental 24-Hour Blood Glucose Profiles,” The Journal of Clinical Endocrinology & Metabolism, vol. 104, no. 1. The Endocrine Society, pp. 111–117, Sep. 24, 2018. doi: 10.1210/jc.2018-01308.

[56] J. Kim, G.M. Saidel, and M.E. Cabrera, "Multi-scale computational model of fuel homeostasis during exercise: Effect of hormonal control," Annals of Biomedical Engineering, vol. 35, (no. 1), pp. 69-90, Jan 2007.

[57] A. Roy and R.S. Parker, "Dynamic modeling of exercise effects on plasma glucose and insulin levels," Journal of diabetes science and technology, vol. 1, (no. 3), pp. 338-347, 2007.

[58] Abu-Rmileh A, Garcia-Gabin W, Zambrano D (2010) A robust sliding mode con-

troller with internal model for closed-loop artificial pancreas. Med Biol Eng Comput 48(12):1191–1201. doi:10.1007/s11517-010-0665-3

[59] Daskalaki E, Norgaard K, Zuger T, Prountzou A, Diem P, Mougiakakou S (2013) An early warning system for hypoglycemic/hyperglycemic events based on fusion of adaptive prediction models. J Diabetes Sci Technol 7(3):689–698

[60] W. L. Clarke, D. Cox, L. A. Gonder-Frederick, W. Carter, and S. L. Pohl, "Evaluating Clinical Accuracy of Systems for Self-Monitoring of Blood Glucose," Diabetes Care, vol. 10, no. 5, pp. 622–628, Sep. 1987, doi: https://doi.org/10.2337/diacare.10.5.622.

[61] "sklearn.linear_model.LinearRegression scikit-learn 0.22 documentation," Scikit-learn.org, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression

[62] "sklearn.linear_model.HuberRegressor" scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.HuberRegressor.html#sklearn.linear_model.HuberRegressor (accessed Apr. 21, 2024).

[63] H. Li and T. Qiu, "Continuous Manufacturing Process Sequential Prediction using Temporal Convolutional Network," Computer-aided chemical engineering/Computer aided chemical engineering, pp. 1789–1794, Jan. 2022, doi: https://doi.org/10.1016/b978-0-323-85159-6.50298-0.

[64] W. L. Haskell et al., "Physical activity and public health: Updated recommendation for adults from the American College of Sports Medicine and the American Heart Association," Medicine & Science in Sports & Exercise, vol. 39, no. 8, pp. 1423-1434, 2007.

[65] Błaszczykowski, Piotr, et al. "Assessing the Accuracy of Smartwatch-Based Heart Rate Variability for Stress and Affect Monitoring: A Systematic Review and Meta-Analysis." Frontiers in Physiology 14 (2023): 779.

[66] O'Connor, Stephen, et al. "Accuracy and consistency of smartwatch-based step counting: A systematic review and meta-analysis." Medicine & Science in Sports & Exercise 53.1 (2021): 139-153.

[67] Chen, Jian, et al. "The Accuracy of Smartwatch-Based Heart Rate Monitoring During Exercise: A Systematic Review and Meta-Analysis." Sports Medicine 50.12 (2020): 2339-2352.

[68] Krawczyk, Marcin, et al. "Challenges and Recommendations for Wearable Devices in Digital Health: Data Quality, Interoperability, Health Equity, Fairness." JAMA Network

Open 6.8 (2023): e233859.

[69] Kriz, J., Tkac, M., & Preiss, J. Machine Learning Models for Blood Glucose Level Prediction in Patients With Diabetes Mellitus: Systematic Review and Network Meta-Analysis. Diabetes Care, 44(10), 2280-2292. 2021. DOI: //doi.org/10.2196%2F47833

[70] Steil, G. M., & Wolfsheimer, K. Continuous glucose monitoring: sensor characteristics and clinical applications. Diabetes Technology & Therapeutics, 3(4), 323-333. 2001. DOI: //doi.org/10.4093%2Fdmj.2019.0121

[71] El-Darrat, R., & Gadge, M. B. Personalized medicine and diabetes: the individualization of therapeutics and treatment monitoring. Clinical Therapeutics, 37(12), 1550-1559. 2015 DOI: https://doi.org/10.1016/j.clinthera.2015.09.004

[72] A. Casolaro, V. Capone, G. Iannuzzo, and F. Camastra, "Deep Learning for Time Series Forecasting: Advances and Open Problems," Information, vol. 14, no. 11. MDPI AG, p. 598, Nov. 04, 2023. doi: 10.3390/info14110598.

[73] D. Fuller et al., "Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review," JMIR mHealth and uHealth, vol. 8, no. 9, p. e18694, 2020, doi: https://doi.org/10.2196/18694.

[74] D. John, A. Morton, D. Arguello, K. Lyden, and D. Bassett, "'What Is a Step?' Differences in How a Step Is Detected among Three Popular Activity Monitors That Have Impacted Physical Activity Research," Sensors, vol. 18, no. 4, p. 1206, Apr. 2018, doi: https://doi.org/10.3390/s18041206.

[75] Miriam, "miriamkw/GluPredKit," GitHub, https://github.com/miriamkw/GluPredKit (accessed Apr. 20, 2024).

[76] A. Hiroshi (1970). "A new method of interpolation and smooth curve fitting based on local procedures" (PDF). Journal of the ACM. 17: 589–602. (accessed Apr. 20, 2024).

[77] M. Hosni, J.M. Carrillo-de-Gea, A. Idri, J.L. Fernández-Alemán, J.A. García-Berná. Using ensemble classification methods in lung cancer disease 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), pp. 1367-1370, 10.1109/EMBC.2019.8857435

[78] J.L. Fernández-Alemán, J.M. Carrillo-de-Gea, M. Hosni, A. Idri, G. García-Mateos. Homogeneous and heterogeneous ensemble classification methods in diabetes disease: a review. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), pp. 3956-3959, 10.1109/EMBC.2019.8856341

[79] G. Seni, J.F. Elder. Ensemble methods in data mining: improving accuracy through combining predictions, synthesis lectures on data mining and knowledge discovery, 2 (2010), pp. 1-126, 10.2200/S00240ED1V01Y200912DMK002

[80] M. Hosni, I. Abnane, A. Idri, J.M. Carrillo de Gea, J.L. Fernández Alemán. Reviewing ensemble classification methods in breast cancer. Comput. Methods Progr. Biomed., 177 (2019), pp. 89-112, 10.1016/j.cmpb.2019.05.019

[81] M. Hosni, J.M. Carrillo de Gea, A. Idri, M. El Bajta, J.L. Fernández Alemán, G. García-Mateos, I. Abnane. A systematic mapping study for ensemble classification methods in cardiovascular disease. Artif. Intell. Rev. (2020), 10.1007/s10462-020-09914-6

[82] G. Brown, J. Wyatt, R. Harris, X. Yao. Diversity creation methods: a survey and categorisation. Inf. Fusion, 6 (2005), pp. 5-20, 10.1016/j.inffus.2004.04.004

[83] Plis, K.; Bunescu, R.; Marling, C.; Shubrook, J.; Schwartz, F. A machine learning approach to predicting blood glucose levels for diabetes management. In Proceedings of the Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–28 July 2014

[84] Munoz-Organero M., "Deep Physiological Model for Blood Glucose Prediction in T1DM Patients," Sensors (Basel), vol. 20, no. 14, p. 3896, Jul. 2020. [Online]. Available: https://doi.org/10.3390/s20143896

[85] Mirshekarian, S.; Bunescu, R.; Marling, C.; Schwartz, F. Using LSTMs to learn physiological models of blood glucose behavior. In Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, South Korea, 11–15 July 2017; pp. 2887–2891.

[86] Gu, W.; Zhou, Z.; Zhou, Y.; He, M.; Zou, H.; Zhang, L. Predicting Blood Glucose Dynamics with Multi-time-series Deep Learning. In Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, Delft, Netherlands, 6–8 November 2017; p. 55.

[87] A. Z. Woldaregay et al., “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, Jul. 2019, doi: https://doi.org/10.1016/j.artmed.2019.07.007.

[88] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” Neural Networks, vol. 2, no. 5, pp. 359–366, July 1989

[89] A. Pfützner, D. C. Klonoff, S. Pardo, and J. L. Parkes, “Technical Aspects of the Parkes

Error Grid," Journal of Diabetes Science and Technology, vol. 7, no. 5, pp. 1275–1281, Sep. 2013, doi: https://doi.org/10.1177/193229681300700517.

[90] A. Facchinetti et al., "A new index to optimally design and compare continuous glucose monitoring glucose prediction algorithms," Diabetes Technol. Therapeutics, vol. 13, no. 2, pp. 111–119, 2011.

[91] C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration," IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1550–1560, Jun. 2012.

[92] J. Xie and Q. Wang, "Benchmarking machine learning algorithms on blood glucose prediction for Type 1 Diabetes in comparison with classical time-series models," IEEE Transactions on Biomedical Engineering, pp. 1–1, 2020, doi: https://doi.org/10.1109/tbme.2020.2975959.

[93] L. Syafaah, S. Basuki, D. Setiawan, A. Faruq, and M. Hery Purnomo, "Diabetes prediction based on discrete and continuous mean amplitude of glycemic excursions using machine learning," Bulletin of Electrical Engineering and Informatics, vol. 9, no. 6, pp. 2619–2629, Aug. 2020, doi: https://doi.org/10.11591/eei.v9i6.2387.

[94] G. Cappon, M. Vettoretti, G. Sparacino, S. D. Favero, and A. Facchinetti, "ReplayBG: A Digital Twin-Based Methodology to Identify a Personalized Model From Type 1 Diabetes Data and Simulate Glucose Concentrations to Assess Alternative Therapies," IEEE transactions on bio-medical engineering, vol. 70, no. 11, pp. 3227–3238, Nov. 2023, doi: https://doi.org/10.1109/TBME.2023.3286856.

[95] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," Neural Comput., vol. 8, no. 7, pp. 1341–1390, 1996