# Physical Activity Integration in Blood Glucose Level Prediction: Different Levels of Data Fusion

Hoda Nemat, Heydar Khadem, Jackie Elliott, and Mohammed Benaissa, *Senior Member, IEEE*

***Abstract*— Blood glucose level (BGL) prediction contributes to more effective management of diabetes. Physical activity (PA), which affects BGL, is a crucial factor in diabetes management. Due to the erratic nature of PA's impact on BGL inter- and intra-patients, deploying PA in BGL prediction is challenging. Hence, it is crucial to discover optimal approaches for utilising PA to improve the performance of BGL prediction. This work contributes to this gap by proposing several PA-informed BGL prediction models. Different approaches are developed to extract information from PA data and integrate this information with BGL data at signal, feature, and decision levels. For signal-level fusion, different automatically-recorded PA data are fused with BGL data. Also, three feature engineering approaches are developed for feature-level fusion: subjective assessments of PA, objective assessments of PA, and statistics of PA. Furthermore, in decision-level fusion, ensemble learning is used to combine predictions from models trained with different inputs. Then, a comparative investigation is performed between the developed PA-informed approaches and the no-fusion approach, as well as between themselves. The analyses are performed on the publicly available Ohio dataset with rigorous evaluation. The results show that among the developed approaches, fusing heart rate data at the signal-level and PA intensity categories at the feature-level with BGL data are effective ways of deploying PA in BGL prediction.**

***Index Terms*— Data fusion, Deep learning, Diabetes management, Ensemble learning, Time series forecasting.**

## I. Introduction

Type 1 diabetes (T1D), a metabolic disorder with different complications, is a significant global cause of morbidity and mortality [1]. Adequate T1D management reduces the risk of complications associated with the disease [2]. Management of T1D entails maintaining Blood glucose levels (BGL) within a target range [1]. Physical activity (PA) is a determinant of insulin sensitivity and an important factor in T1D management [3]–[5]. Due to insufficient explicit knowledge of how exactly PA impacts BGL, optimal diabetes management is hindered in the presence of PA, and it is difficult for clinicians to provide patients with specific advice concerning PA [6]. Part of the complexity arises because it has recently been shown that BGL can vary significantly for individuals during and after exercise from one day to another, even for the same type and duration of exercise performed at the same time of day and after consuming similar meals[7]. Hence, although regular exercise is beneficial for T1D patients as it helps reduce cardiovascular disease risk, maintaining normoglycaemia is challenging. Indeed, many people with T1D avoid exercise so as to not increase the chances of hyperglycaemia or hypoglycaemia events before or after exercise [8]–[10].

Accurate BGL prediction models can facilitate more effective glycaemic control. Such models predict future BGLs from current and past data and provide early warnings about possible abnormal glycaemic events [11]–[13]. Predicting BGL accurately remains challenging, and any improvements are highly valued. In this regard, recent studies deployed different advanced artificial intelligence techniques, including deep learning [14]–[20], ensemble learning [21]–[27], transfer learning [17], [28]–[31], and causal inference [32] for further improvements in BGL prediction performance. Some studies by deploying affecting variables on BGL, including carbohydrate intake, injected insulin, and PA, tried to make further improvement in BGL prediction performance [17], [33]–[35]. Although it is postulated that these variables are effective variables on BGL, according to the literature, incorporating them in the BGL prediction tasks may not improve the prediction's performance, and there is no conclusive decision regarding the optimal input for the BGL prediction task [36]–[38]. Hence, it is essential to discover effective approaches to incorporate each variable into BGL prediction. Among these influencing factors, handling PA in BGL prediction is particularly challenging due to its significantly varying effect on BGL. Therefore, it is crucial to discover optimal approaches for leveraging PA in BGL prediction.

Recently, several types of research have been developed to process and fuse different information using deep learning [39]–[43]. In diabetes, limited studies have investigated some fusion approaches to improve the BGL prediction performance. Dudukcu et al. [44], by developing three neural networks and fusing their outputs at the decision level, improved the BGLP prediction for the prediction horizon of 30 minutes. Also, Khadem et al. [26], to improve the performance of BGL prediction, proposed a lag fusion network using meta-learning analysis. In their work, MLP and LSTM models were trained four times over histories of 30, 60, 90, and 120 minutes, and then their information was fused. These studies, however, used only BGL data to develop their models, and there is still a demand to investigate effective approaches for fusing the information of other affecting variables in BGL prediction. This work aims to discover optimal approaches for leveraging PA in BGL prediction by developing and comparing a wide range of approaches for different levels of data fusion.

This work proposes various approaches for extracting different types of information from PA data and fusing this PA-driven information with BGL. Different levels of fusion, including signal-level, feature-level, and decision-level fusion, are investigated to find effective ways of integrating PA into BGL prediction. To do so, univariate and multivariate LSTM models are generated to predict BGL without and with different PA fusion approaches. For signal-level data fusion, combinations of raw PA data directly collected from wristbands are examined. Also, for feature-level data fusion, three feature engineering approaches are developed; subjective assessments of PA, objective assessments of PA, and statistics of PA. Lastly, in decision-level data fusion, ensemble learning is used to

This work is not associated with funding Agency. (Corresponding author: Hoda Nemat.)

Hoda Nemat and Mohammed Benaissa are with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, S1 4DE, U.K. (e-mail: hoda.nemat@sheffield.ac.uk; m.benaissa@sheffield.ac.uk).

Heydar Khadem is with the Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, S1 4DE, U.K., also with the Department of Computer Science, University of Manchester, M13 9PL Manchester, U.K., and also with the Artificial Intelligence Machine Learning Team, EC4A 1HP KultraLab, London, U.K.. (e-mail: h.khadem@sheffield.ac.uk).

Jackie Elliott is with the Department of Oncology and Metabolism, University of Sheffield, Sheffield S10 2RX, U.K., and also with the Sheffield Teaching Hospitals, Diabetes and Endocrine Centre, Northern General Hospital, Sheffield S5 7AU, U.K. (e-mail: j.elliott@sheffield.ac.uk).

combine predictions from multiple models to make final predictions, similar to the techniques proposed in our previous conference paper [45]. The analyses are performed using real T1D subjects from the publicly available Ohio dataset [46], and the performance of different PA-informed prediction models is evaluated using regression-based and clinical-based metrics. Moreover, rigorous statistical analysis is performed to have a valid and conclusive deduction.

## II. Related Works

Several investigations have been performed examining PA in T1D management. As part of our review, we categorise the relevant PA-involved works in T1D into three main areas: detection, classification, or description of PA [47]–[50], investigation of the impact of PA on T1D management [7], [51], [52], and inclusion of PA in adverse glycaemic events detection and BGL prediction [5], [53]–[55]. The following paragraphs provide a brief overview of these works.

In relation to the analysis of PA itself, in a study by Cho et al. [47], accelerometer, heart rate (HR), and continuous glucose monitoring (CGM) data were used to detect and classify the type and intensity of PA by developing random forest models. Also, Cescon et al. [48] detected and classified PA based on its intensity by developing deep learning models using accelerometer data collected from wristbands in free-living conditions. Moreover, Dénes-Fazakas et al. [49], by investigating different machine learning (ML) models, detected the presence of physical activity from CGM and HR data. Also, Ozaslan et al. [50] proposed a physiological model for activity on board using step counts.

ML approaches have been used in some studies to investigate the impact of PA on managing T1D. By analysing data from 37 T1D patients using linear regression models, Ozaslan et al. [51] concluded that PA can have immediate and delayed effects on BGL. Also, they found that there is a significant relationship between PA and BGL after an evening meal, suggesting that measuring PA may be helpful for guiding meals. Also, Ozaslan et al. [52] proposed an insulin dosing system by adding PA information, which significantly decreased time spent in hypoglycaemia and increased time spent in normoglycaemia. In their study, Tyler et al. [7] collected a dataset of highly-controlled exercise sessions and investigated several ML models to quantify the effect of physical activity on BGL. They developed an adaptive, personalized ML model to predict exercise-related BGL changes. Moreover, in some studies, to cope with artifacts and disturbances and to develop a more comprehensive overall indicator of PA, the incorporation of multiple physiological signals for PA has been considered for better adjustment of insulin delivery [56]–[58].

Some studies have been performed to predict BGL or glycaemic events by including PA. Xie and Wang [5] developed a glucose dynamics model by considering PA. By entering the PA, they proposed a non-linear autoregressive moving average model with exogenous inputs. To train and evaluate the model, they used in silico data from the UVa/Padova simulator. They observed that during and two hours after exercise, the nonlinear and linear models with PA made a better prediction for BGL in a prediction horizon of 30 minutes than the linear model without PA. Also, Bertachi et al. [54] investigated the possibility of nocturnal hypoglycaemia prediction in T1D by incorporating PA. They analysed the data of CGM sensors and physical activity trackers collected in 12 weeks from 10 people with T1D. They applied MLP and SVM models for binary classification. They concluded it was feasible to predict nocturnal hypoglycaemia from CGM and activity data using ML approaches. Moreover, Hobbs et al. [53] developed a glycemic model by considering some terms indicating different effects of PA on metabolism. They showed their model outperformed the prediction model using only BGL.

Although several studies have focused on PA in T1D management, to the best of our knowledge, no study has investigated different levels of PA fusion in BGL prediction. Hence, This work contributes to this gap by developing several PA-informed methods for BGL prediction.

## III. Dataset

To accurately evaluate the developed methods, we used the Ohio T1D dataset released in 2018 for the BGL prediction challenge [46], and is the most commonly used publicly available clinical dataset in the literature [59]. The dataset contained data from physiological sensors and self-reported life events of six individuals with T1D. The participants were four females and two males, aged between 40 and 60. Each participant had two distinct XML files for training and testing sets. The total data for each patient was eight weeks' worth, of which the last 10 days were allocated to the testing set and the rest used for the training set.

BGL data was collected with a 5-minute aggregation using a Medtronic Enlite CGM sensor. PA data was automatically recorded as heart rate (HR), step count (SC), galvanic skin response (GSR), and skin temperature (ST) using a Basis Peak band with 5-minute aggregation. Also, patients reported times and duration of sleep, work, and exercise. An individual's subjective assessment of physical effort was measured for work and exercise on a scale from one to 10, with 10 indicating the highest level of PA.

The number of data points for BGL and automatic-recorded PA data is provided in Table I. Also, Table II shows the count of self-reported data related to PA with patients' subjective assessment of intensity level. Furthermore, additional details regarding the dataset, including more information on sensors and devices, can be found in [46].

TABLE I: The number of data points for blood glucose and automatic-recorded physical activity in training and testing sets related to the contributors in the Ohio dataset.

| PID | Training data points | | Testing data points | |
|---|---|---|---|---|
| | BGL | PA band | BGL | PA band |
| 559 | 10796 | 11979 | 2514 | 2633 |
| 563 | 12124 | 11966 | 2570 | 2706 |
| 570 | 10982 | 12328 | 2745 | 2720 |
| 575 | 11866 | 12446 | 2590 | 2698 |
| 588 | 12640 | 12980 | 2791 | 2620 |
| 591 | 10847 | 12276 | 2760 | 2668 |

Note. PID: patient identity; BGL: blood glucose level; PA: physical activity.

In the present work, CGM data and automatic-recorded and self-reported data related to PA were used. Figure 1 shows BGL and PA-related data for a duration of 24 hours of training data for the data contributor with PID 559.

## IV. Methods

### A. Preprocessing

First, the missing data had to be dealt with in the preprocessing phase. To do so, the missing BGL, HR, GSR, and ST data were imputed using a linear approach, with interpolation and extrapolation techniques used in the training and testing sets, respectively. Also, missing SC data were filled with zero values for non-reported data timestamps. Aligning BGL and PA data was the next preprocessing step. Additionally, since CGM sensors and activity bands were worn at different times, some data were unavailable at the beginning or end of each set. To have more reliable data, these timestamps were excluded from the analysis. Moreover, the time series forecasting task of BGL prediction was recast to a supervised learning task. To do

TABLE II: The number and patients' subjective assessment of intensity levels of physical activity data.

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

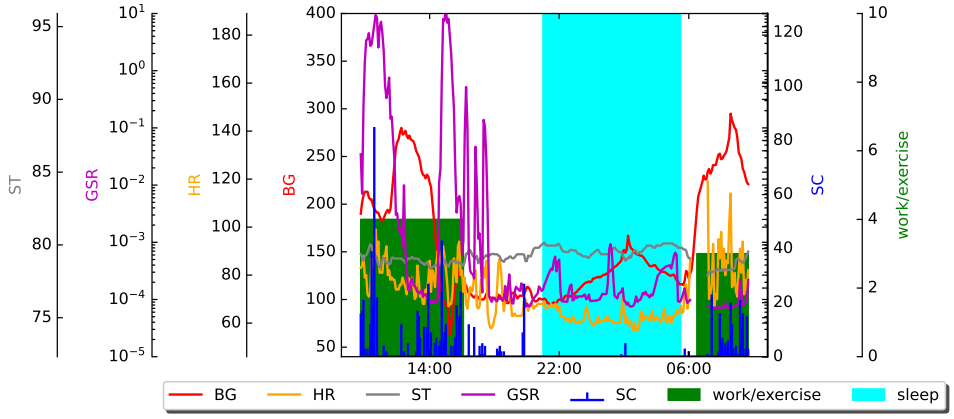



Fig. 1: A presentation of blood glucose level and physical activity-related data for PID 559 for a duration of 24 hours. Note. BG: blood glucose; HR: heart rate; GSR: galvanic skin response; ST: skin temperature; SC: step count.

so, the time series data were sampled, with lag observations serving as input and future observations serving as output. Then, using a 5-minute rolling window, history samples with 60-minute lengths were assigned as input, and future samples with 30-minute and 60-minute lengths were assigned as output. Finally, the scaling process was applied to the input samples for each variable based on its minimum and maximum values over the entire training set.

### B. Prediction models

LSTM networks, a type of recurrent neural network, are capable of predicting BGL [15], [24], [45], [60]. An LSTM model recently developed by our team, as described in [23], [32], was used for the BGL prediction task in the current work. The vanilla LSTM network used had three layers: a 200-unit LSTM layer, a 100-unit dense layer, and an output layer of one unit dense layer. The initialiser, activation function, optimiser, and loss function were chosen as He uniform, ReLU, Adam, and mean square error, respectively. Also, the epoch size was set at 200 and the batch size at 32. The univariate LSTM model was used for prediction using only BGL, and it is called the no-fusion approach. Furthermore, according to the fusion approach, both univariate and multivariate LSTM models were used for PA-informed approaches.

It is worth noting that to develop the LSTM networks, two categories of parameters needed to be determined: parameters related to the network's structure, including the number of hidden layers, the number of units in each layer, the initialiser, and the activation function, and parameters related to compilation, including batch size, learning rate, and the optimiser. To do so, we partitioned the training set into the first 80% for training and the following 20% for validation purposes. The hyperparameters were then optimised by contributing to the lowest average RMSE. We started with one hidden layer with low units. By achieving good predictability capacity on the evaluation set, we kept our LSTM model shallow with just one hidden layer. Also, due to computational limitations, we used random searches to determine the initialiser, activation function, and optimiser. More details about the model architecture and optimization for univariate and multivariate models can be found in [23] and [32].

### C. PA fusion

From a data fusion perspective, ML approaches for data fusion are categorised into three levels: signal fusion, feature fusion, and decision fusion [61], [62]. In this work, different kinds of information from PA at different levels were fused with BGL data. The performance of the BGL prediction for different fusion information/levels was investigated and compared with the prediction model without PA information. In the following, the approaches for each level of data fusion are described.

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

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

*a) Subjective PA categorisation:* Considering that PA is defined as any motion generated by skeletal muscle that increases energy expenditure, it can be categorised as sedentary, light, moderate, and vigorous in terms of relative effort and expenditure of energy [63]. In the first feature extraction technique, self-reported data related to PA were deployed as PA features. To do so, subjective assessments of participants for physical exertion, which were scaled from one to 10,

were categorised into three different intensity levels. In detail, data reported with scales of one, two, and three were assigned to the light category; data reported with scales of four and five were allocated to the moderate category; and data reported with a scale of six or more were categorised as vigorous. Also, non-reported timestamps were assumed to be inactive and assigned to the sedentary category. Moreover, sleep data was assigned to a separate category. Hence, five categories related to different levels of PA intensities, including sleep, sedentary, light, moderate, and vigorous, were used as subjective PA features. These features were then employed as input along with the BGL data for the multivariate LSTM prediction model. In short, this approach is called BG+SPA.

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

*c) Statistics of PA data:* In this feature category, the statistics of the automatic-recorded PA data, which have been shown to be effective in the BGL prediction in the literature [45], [64], were used for the fusion with BGL data. To do so, PA statistics, including the mean and standard deviation, were calculated for all the automatic-recorded PA data and added to the corresponding history of BGL data. This was fed as the input of the univariate LSTM model. It is called the BG+StPA approach.

*3) Decision-level PA fusion:* The highest level of data fusion is decision fusion, which combines information that has already generated some decisions for a given task. To examine this level of data fusion, a method employing stacked ensemble learning [65] was developed based on the idea we proposed in our conference paper [45]. The stacked regression consists of multiple models serving as base-learners and a meta-learner fed by the outputs of the base-learners. In this work, instead of using different models as base-learners, the univariate LSTM model was trained twice, once using BGL data and once using PA data. Thus, at the first level of learning, primary decisions were generated separately using BGL and PA data. The decisions of the first layer were then stacked and used as input for the meta-learner, which was chosen as a linear regression model to provide the final prediction. Accordingly, deploying the concept of ensemble learning, PA information was fused with the BGL in a decision-level approach.

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

### D. Evaluation criteria

In this study, the performance of BGL prediction using different approaches for data fusion of PA was evaluated and compared for two prediction horizons of 30 and 60 minutes. The evaluation was performed based on regression-wised and clinical-wised criteria. The regression-wised criteria included root mean square error (RMSE) and mean absolute error (MAE). The clinical-wised criteria included the Matthews correlation coefficient (MCC) and surveillance error (SE). Definitions and equations regarding these metrics can be found in our recent articles [23], [26], [32], [62], [66]–[69].

### E. Statistical analyses

The performance of BGL prediction using various data fusion approaches was also statistically evaluated and compared over data contributors. Approaches for each level of PA fusion and the no-fusion approach were compared pair-wisely based on the recommended statistical tests in [70]. To do so, using a Friedman test [71], it was determined if there was a significant difference in the performance of BGL prediction between at least two approaches. Next, the Post-hoc Nemenyi test [72] was performed for pair-wise comparisons to determine which approaches performed significantly differently in a pair-wise fashion, with a significance level of 5%. Furthermore, the results of the post-hoc test were depicted by a critical difference (CD) diagram [70]. These analyses were then also performed between effective PA fusion approaches of each fusion level.

## V. Results and Discussion

This section presents the evaluation results of different PA fusion approaches along with rigorous statistical analyses for the two prediction horizons of 30 and 60 minutes. It is worth noting that since LSTM models rely on random initialisation, their performance was evaluated ten times, and the mean and standard deviation are reported for evaluation metrics.

### A. No-fusion prediction

Table III presents the evaluation results of the BGL prediction using the no-fusion approach, in which BGL data was used as the only input for prediction horizons of 30 and 60 minutes.

### B. PA-fused prediction

*1) Signal-level PA fusion:* Table IV shows the results of evaluating the BGL prediction models that use BGL data fused with different signal-level information from PA to make predictions 30 and 60 minutes in advance.

To have a pair-wise comparison between the no-fusion approach and signal-level PA fusion approaches, first, the Friedman test was performed for both prediction horizons and all evaluation metrics. According to Table V, there is sufficient evidence to be inferred that at least two approaches may perform differently for the BGL prediction. Therefore, in the next step, the post-hoc Nemenyi test was performed for pair-wise comparisons to determine which PA fusion approaches performed significantly differently. The results of the Nemenyi tests based on each evaluation metric are graphically represented as CD diagrams where horizontal lines link approaches with similar performances at a significance level of 5%. Then, to have an overview, CD diagrams according to the average ranking over all evaluation criteria were generated for each prediction horizon of 30 and 60 minutes. To be concise, individual CD diagrams related to each metric are presented in the Appendix section (Figure 6), and CD diagrams based on the average over all metrics are presented in Figure 2.

Considering Figures 2a and 2b, which show the ranking for prediction horizon of 30 and 60 minutes, respectively, it can be concluded that the BG+HR approach was the best approach among signal-level PA fusion approaches and statistically significantly outperformed the no-fusion approach. Considering Tables III and IV and Figure 6, it can be concluded that among developed signal-level fusion approaches, the BG+HR approach improved the average SE by 5.296% for the prediction horizon of 30 minutes. Also, for the prediction horizon of 60 minutes, this fusion approach improved RMSE, MAE, MCC, and SE by 4.466%, 6.403%, 7.479%, and 7.539%, respectively.

TABLE III: Evaluation results of the BGL prediction using no-fusion approach for prediction horizons of 30 and 60 minutes.

| PID | PH: 30 min | | | | PH: 60 min | | | |
|---|---|---|---|---|---|---|---|---|
| | RMSE | MAE | MCC | SE | RMSE | MAE | MCC | SE |
| 559 | 19.854 ± 0.183 | 13.948 ± 0.222 | 0.787 ± 0.006 | 0.202 ± 0.004 | 35.070 ± 0.113 | 25.878 ± 0.186 | 0.618 ± 0.013 | 0.350 ± 0.006 |
| 563 | 18.800 ± 0.088 | 13.089 ± 0.077 | 0.770 ± 0.002 | 0.182 ± 0.001 | 34.156 ± 1.851 | 25.462 ± 1.782 | 0.459 ± 0.061 | 0.347 ± 0.025 |
| 570 | 23.504 ± 0.607 | 17.515 ± 0.628 | 0.820 ± 0.003 | 0.158 ± 0.004 | 29.007 ± 0.368 | 21.111 ± 0.271 | 0.790 ± 0.008 | 0.197 ± 0.003 |
| 575 | 24.075 ± 0.357 | 15.504 ± 0.521 | 0.730 ± 0.004 | 0.241 ± 0.012 | 37.858 ± 0.279 | 27.827 ± 1.483 | 0.521 ± 0.019 | 0.425 ± 0.033 |
| 588 | 18.881 ± 0.076 | 13.603 ± 0.053 | 0.738 ± 0.005 | 0.181 ± 0.002 | 37.783 ± 3.989 | 28.463 ± 3.451 | 0.430 ± 0.067 | 0.382 ± 0.049 |
| 591 | 22.683 ± 0.149 | 16.465 ± 0.103 | 0.639 ± 0.009 | 0.284 ± 0.002 | 37.851 ± 0.795 | 29.952 ± 0.796 | 0.378 ± 0.008 | 0.471 ± 0.011 |
| Avg | 21.299 ± 0.243 | 15.021 ± 0.267 | 0.747 ± 0.005 | 0.208 ± 0.004 | 35.288 ± 1.233 | 26.449 ± 1.328 | 0.533 ± 0.029 | 0.362 ± 0.021 |

Note. PID: patient identity; PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error.

TABLE IV: Evaluation results of the BGL prediction using signal-level physical activity fusion approaches for prediction horizons of 30 and 60 minutes.

| PID | Input | PH: 30 min | | | | PH: 60 min | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RMSE | MAE | MCC | SE | RMSE | MAE | MCC | SE |
| 559 | BG+HR | 19.980 ± 0.053 | 13.857 ± 0.130 | 0.813 ± 0.005 | 0.185 ± 0.004 | 35.169 ± 0.333 | 25.619 ± 0.301 | 0.632 ± 0.012 | 0.333 ± 0.011 |
| | BG+HRSC | 23.691 ± 0.609 | 16.067 ± 0.258 | 0.782 ± 0.009 | 0.214 ± 0.003 | 39.131 ± 1.062 | 28.424 ± 0.884 | 0.609 ± 0.028 | 0.369 ± 0.020 |
| | BG+Band | 23.411 ± 0.523 | 16.141 ± 0.231 | 0.765 ± 0.005 | 0.216 ± 0.004 | 40.323 ± 1.131 | 29.019 ± 0.712 | 0.589 ± 0.010 | 0.382 ± 0.012 |
| 563 | BG+HR | 18.942 ± 0.139 | 13.213 ± 0.122 | 0.766 ± 0.006 | 0.183 ± 0.002 | 31.398 ± 1.190 | 23.039 ± 1.171 | 0.543 ± 0.036 | 0.315 ± 0.020 |
| | BG+HRSC | 19.256 ± 0.187 | 13.475 ± 0.184 | 0.771 ± 0.006 | 0.186 ± 0.002 | 31.674 ± 0.198 | 23.384 ± 0.252 | 0.552 ± 0.009 | 0.314 ± 0.003 |
| | BG+Band | 19.408 ± 0.155 | 13.665 ± 0.098 | 0.772 ± 0.008 | 0.188 ± 0.002 | 32.042 ± 0.390 | 23.311 ± 0.583 | 0.543 ± 0.022 | 0.316 ± 0.013 |
| 570 | BG+HR | 16.550 ± 0.136 | 11.530 ± 0.108 | 0.871 ± 0.005 | 0.112 ± 0.002 | 28.584 ± 0.535 | 20.792 ± 0.453 | 0.785 ± 0.006 | 0.195 ± 0.004 |
| | BG+HRSC | 16.886 ± 0.520 | 11.669 ± 0.318 | 0.866 ± 0.005 | 0.113 ± 0.005 | 28.540 ± 0.323 | 20.775 ± 0.231 | 0.782 ± 0.008 | 0.195 ± 0.003 |
| | BG+Band | 17.929 ± 0.362 | 12.352 ± 0.351 | 0.849 ± 0.005 | 0.122 ± 0.004 | 29.655 ± 0.468 | 21.508 ± 0.382 | 0.753 ± 0.003 | 0.209 ± 0.004 |
| 575 | BG+HR | 23.996 ± 0.405 | 15.264 ± 0.254 | 0.742 ± 0.018 | 0.225 ± 0.006 | 37.712 ± 0.236 | 26.430 ± 0.167 | 0.527 ± 0.014 | 0.395 ± 0.004 |
| | BG+HRSC | 24.317 ± 0.133 | 15.552 ± 0.232 | 0.740 ± 0.015 | 0.229 ± 0.005 | 38.549 ± 0.573 | 27.368 ± 0.449 | 0.518 ± 0.013 | 0.405 ± 0.009 |
| | BG+Band | 24.449 ± 0.188 | 15.597 ± 0.164 | 0.742 ± 0.011 | 0.230 ± 0.002 | 39.246 ± 0.387 | 28.017 ± 0.330 | 0.519 ± 0.014 | 0.415 ± 0.006 |
| 588 | BG+HR | 18.916 ± 0.109 | 13.858 ± 0.294 | 0.750 ± 0.013 | 0.187 ± 0.008 | 32.305 ± 0.791 | 23.456 ± 0.400 | 0.556 ± 0.018 | 0.307 ± 0.003 |
| | BG+HRSC | 19.320 ± 0.341 | 13.984 ± 0.298 | 0.736 ± 0.017 | 0.186 ± 0.003 | 32.422 ± 0.403 | 23.605 ± 0.327 | 0.556 ± 0.013 | 0.305 ± 0.005 |
| | BG+Band | 19.483 ± 0.185 | 14.060 ± 0.172 | 0.730 ± 0.003 | 0.186 ± 0.001 | 33.430 ± 0.508 | 24.438 ± 0.344 | 0.546 ± 0.005 | 0.314 ± 0.004 |
| 591 | BG+HR | 22.644 ± 0.397 | 16.228 ± 0.388 | 0.637 ± 0.009 | 0.280 ± 0.007 | 37.102 ± 0.792 | 29.194 ± 0.976 | 0.392 ± 0.016 | 0.463 ± 0.013 |
| | BG+HRSC | 22.807 ± 0.391 | 16.380 ± 0.345 | 0.641 ± 0.011 | 0.282 ± 0.006 | 36.995 ± 1.171 | 28.632 ± 1.309 | 0.411 ± 0.017 | 0.458 ± 0.020 |
| | BG+Band | 22.948 ± 0.268 | 16.398 ± 0.226 | 0.653 ± 0.010 | 0.281 ± 0.004 | 36.833 ± 0.906 | 28.844 ± 1.232 | 0.428 ± 0.033 | 0.462 ± 0.020 |
| Avg | BG+HR | 20.171 ± 0.207 | 13.992 ± 0.216 | 0.763 ± 0.009 | 0.195 ± 0.005 | 33.712 ± 0.646 | 24.755 ± 0.578 | 0.572 ± 0.017 | 0.335 ± 0.009 |
| | BG+HRSC | 21.046 ± 0.364 | 14.521 ± 0.273 | 0.756 ± 0.011 | 0.202 ± 0.004 | 34.552 ± 0.622 | 25.365 ± 0.575 | 0.571 ± 0.015 | 0.341 ± 0.010 |
| | BG+Band | 21.271 ± 0.280 | 14.702 ± 0.207 | 0.752 ± 0.007 | 0.204 ± 0.003 | 35.255 ± 0.632 | 25.856 ± 0.597 | 0.563 ± 0.014 | 0.350 ± 0.010 |

Note. PID: patient identity; PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error; BG+HR, BG+HRSC, and BG+Band: approaches for the signal-level fusion of physical activity data with blood glucose data.

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

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

Note. PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error.

*2) Feature-level PA fusion:* The evaluation results of BGL prediction 30 and 60 minutes in advance using BGL data fused with different PA-driven features are presented in Table VI.

Table VII shows the p-values of the Friedman test comparing the no-fusion approach and feature-level PA fusion approaches. Whenever the p-value for the Friedman test was significant for any metric, a Nemenyi test was performed and visualised as CD diagrams. Similar to the previous section, CD diagrams related to each metric are shown in the Appendix section (Figure 7), and CD diagrams based on average over all the significant metrics are visualised in Figure 3.

According to Figures 3a and 3b, it can be inferred that the BG+SPA approach was the best among feature-level PA fusion approaches, outperforming the no-fusion approach for both prediction horizons of 30 and 60 minutes. Considering Tables III and VI and Figure 7, it can be inferred that the BG+SPA approach improved the average evaluation metrics of MCC and SE over all patients by 2.921% and 6.063%, respectively, for the prediction horizon of 30 minutes, compared to the no-fusion approach. In addition, the BG+SPA improved the average values of MAE and SE by 6.495% and 8.183%, respectively, compared to the no-fusion approach for the prediction horizon of 60 minutes.

*3) Decision-level PA fusion:* Table VIII displays the evaluation results of BGL prediction models that fuse decision-level information of PA and BGL using ensemble learning for prediction horizons of 30 and 60 minutes.

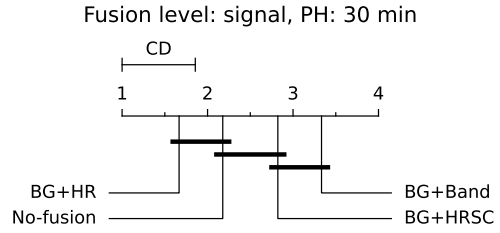
The Friedman test was performed to compare the no-fusion and decision-level PA fusion approaches (Table IX). This was followed by the post-hoc Nemenyi test for significantly different metrics. CD diagrams visualising Nemenyi tests related to these metrics are shown in the Appendix section (Figure 8), and CD diagrams based on average over all the metrics are visualised in Figure 4.

Based on Figures 4a and 4b, it can be concluded that for both

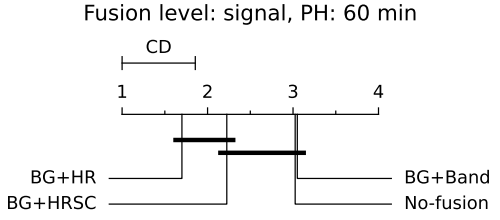
TABLE VI: Evaluation results of the BGL prediction using feature-level physical activity fusion approaches for prediction horizons of 30 and 60 minutes.

| PID | Input | PH: 30 min | | | | PH: 60 min | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RMSE | MAE | MCC | SE | RMSE | MAE | MCC | SE |
| 559 | BG+SPA | 20.285 ± 0.270 | 14.208 ± 0.152 | 0.807 ± 0.006 | 0.189 ± 0.004 | 35.785 ± 0.483 | 26.098 ± 0.246 | 0.615 ± 0.010 | 0.339 ± 0.004 |
| | BG+OPA | 21.394 ± 0.279 | 14.897 ± 0.120 | 0.793 ± 0.009 | 0.198 ± 0.005 | 36.592 ± 0.150 | 26.550 ± 0.144 | 0.628 ± 0.015 | 0.341 ± 0.006 |
| | BG+StPA | 21.011 ± 0.238 | 14.616 ± 0.066 | 0.798 ± 0.008 | 0.201 ± 0.004 | 35.502 ± 0.339 | 25.965 ± 0.294 | 0.609 ± 0.019 | 0.347 ± 0.007 |
| 563 | BG+SPA | 18.881 ± 0.072 | 13.171 ± 0.093 | 0.773 ± 0.005 | 0.183 ± 0.002 | 31.401 ± 0.461 | 22.958 ± 0.589 | 0.554 ± 0.030 | 0.313 ± 0.012 |
| | BG+OPA | 19.015 ± 0.077 | 13.321 ± 0.079 | 0.771 ± 0.006 | 0.184 ± 0.002 | 31.462 ± 0.678 | 23.161 ± 0.577 | 0.552 ± 0.016 | 0.315 ± 0.009 |
| | BG+StPA | 20.508 ± 0.430 | 14.227 ± 0.415 | 0.731 ± 0.018 | 0.201 ± 0.007 | 32.895 ± 0.114 | 23.567 ± 0.321 | 0.509 ± 0.020 | 0.326 ± 0.008 |
| 570 | BG+SPA | 16.404 ± 0.230 | 11.382 ± 0.101 | 0.868 ± 0.005 | 0.112 ± 0.001 | 29.101 ± 1.030 | 21.009 ± 0.740 | 0.772 ± 0.008 | 0.201 ± 0.003 |
| | BG+OPA | 17.023 ± 0.140 | 11.942 ± 0.113 | 0.870 ± 0.008 | 0.115 ± 0.002 | 28.798 ± 0.337 | 21.119 ± 0.247 | 0.785 ± 0.015 | 0.198 ± 0.005 |
| | BG+StPA | 17.053 ± 0.477 | 11.959 ± 0.417 | 0.868 ± 0.004 | 0.114 ± 0.003 | 28.667 ± 0.445 | 20.924 ± 0.414 | 0.786 ± 0.006 | 0.194 ± 0.002 |
| 575 | BG+SPA | 23.913 ± 0.201 | 15.299 ± 0.070 | 0.735 ± 0.004 | 0.223 ± 0.003 | 37.955 ± 0.439 | 26.737 ± 0.229 | 0.499 ± 0.010 | 0.392 ± 0.004 |
| | BG+OPA | 24.605 ± 0.411 | 15.461 ± 0.267 | 0.738 ± 0.006 | 0.226 ± 0.003 | 37.977 ± 0.344 | 26.605 ± 0.252 | 0.521 ± 0.015 | 0.390 ± 0.004 |
| | BG+StPA | 23.520 ± 0.458 | 15.035 ± 0.324 | 0.753 ± 0.012 | 0.229 ± 0.009 | 37.593 ± 0.459 | 26.549 ± 0.696 | 0.532 ± 0.028 | 0.402 ± 0.018 |
| 588 | BG+SPA | 18.530 ± 0.414 | 13.669 ± 0.270 | 0.769 ± 0.006 | 0.181 ± 0.006 | 31.423 ± 0.222 | 22.680 ± 0.180 | 0.597 ± 0.010 | 0.294 ± 0.004 |
| | BG+OPA | 19.018 ± 0.238 | 13.738 ± 0.193 | 0.754 ± 0.009 | 0.182 ± 0.004 | 32.311 ± 0.241 | 23.288 ± 0.179 | 0.579 ± 0.007 | 0.301 ± 0.003 |
| | BG+StPA | 18.627 ± 0.135 | 13.708 ± 0.235 | 0.755 ± 0.005 | 0.181 ± 0.004 | 31.169 ± 0.307 | 22.603 ± 0.307 | 0.557 ± 0.007 | 0.296 ± 0.005 |
| 591 | BG+SPA | 22.595 ± 0.149 | 16.450 ± 0.147 | 0.663 ± 0.002 | 0.284 ± 0.003 | 36.649 ± 1.116 | 28.904 ± 1.150 | 0.446 ± 0.014 | 0.456 ± 0.016 |
| | BG+OPA | 22.858 ± 0.421 | 16.302 ± 0.276 | 0.650 ± 0.004 | 0.277 ± 0.003 | 36.696 ± 0.798 | 28.860 ± 0.885 | 0.389 ± 0.010 | 0.455 ± 0.012 |
| | BG+StPA | 22.276 ± 0.421 | 16.125 ± 0.380 | 0.633 ± 0.020 | 0.276 ± 0.008 | 34.806 ± 0.756 | 26.553 ± 0.811 | 0.436 ± 0.036 | 0.426 ± 0.013 |
| Avg | BG+SPA | 20.101 ± 0.222 | 14.030 ± 0.139 | 0.769 ± 0.005 | 0.195 ± 0.003 | 33.719 ± 0.625 | 24.731 ± 0.522 | 0.580 ± 0.014 | 0.332 ± 0.007 |
| | BG+OPA | 20.652 ± 0.261 | 14.277 ± 0.175 | 0.763 ± 0.007 | 0.197 ± 0.003 | 33.972 ± 0.425 | 24.931 ± 0.381 | 0.576 ± 0.013 | 0.333 ± 0.007 |
| | BG+StPA | 20.499 ± 0.360 | 14.278 ± 0.306 | 0.756 ± 0.011 | 0.200 ± 0.006 | 33.439 ± 0.403 | 24.360 ± 0.474 | 0.572 ± 0.019 | 0.332 ± 0.009 |

Note. PID: patient identity; PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error; BG+SPA, BG+OPA, and BG+StPA: approaches for the feature-level fusion of physical activity data with blood glucose data.




(a)




(b)

Fig. 2: Critical difference diagram showing the comparison of the no-fusion approach and signal-level physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

prediction horizons of 30 and 60 minutes, the BG&HR approach outperformed the no-fusion approach. Considering Tables III and VIII and Figure 8, it can be inferred that the BG&HR approach improved the average evaluation metrics of RMSE and MAE over all patients by 5.265% and 6.684%, for the prediction horizon of 30 minutes, respectively, compared to the no-fusion approach. Moreover,

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

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

Note. PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error.

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

*4) Comparison of different approaches:* As mentioned previously, BG+HR, BG+SPA, and BG&HR approaches outperformed the no-fusion approach for at least one evaluation metric for both prediction horizons. A Friedman test was performed according to all evaluation metrics to compare these approaches. According to the p-values of the Friedman test (Table X), there was a significant difference between at least two PA fusion approaches regarding the MCC and SE evaluation metrics. Hence, the post-hoc Nemenyi test was performed on these metrics for pairwise comparisons. Similarly, CD diagrams visualising the outputs of Nemenyi tests based on each metric are shown in Figure 9 in the Appendix section. Also, CD diagrams based on the average over the two metrics are displayed in Figure 5.

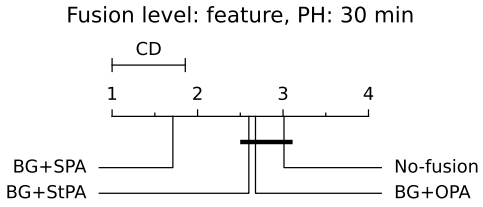
Considering Figure 5, it can be concluded that BG+HR and BG+SPA approaches similarly performed better than the BG&HR approach. Also, it is worth mentioning that based on Tables III, IV, VI, and VIII, PA fusion approaches for PID 570, which based on Table II could be considered as the most active patient, impacted more significantly on BGL prediction performance than other patients.

The results show that the developed approaches for different

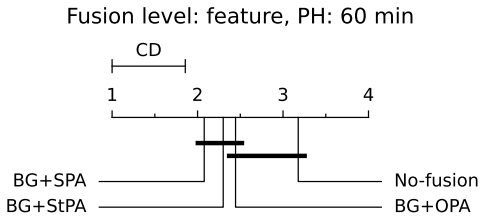
TABLE VIII: Evaluation results of the BGL prediction using decision-level physical activity fusion approaches for prediction horizons of 30 and 60 minutes.

| PID | Input | PH: 30 min | | | | PH: 60 min | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | RMSE | MAE | MCC | SE | RMSE | MAE | MCC | SE |
| 559 | BG&HR | 19.488 ± 0.060 | 13.548 ± 0.053 | 0.791 ± 0.007 | 0.196 ± 0.002 | 34.370 ± 0.287 | 25.534 ± 0.194 | 0.617 ± 0.004 | 0.348 ± 0.003 |
| | BG&HRSC | 19.574 ± 0.065 | 13.631 ± 0.064 | 0.792 ± 0.006 | 0.197 ± 0.002 | 34.854 ± 0.241 | 26.022 ± 0.159 | 0.603 ± 0.009 | 0.354 ± 0.002 |
| | BG&Band | 19.605 ± 0.078 | 13.655 ± 0.057 | 0.794 ± 0.005 | 0.196 ± 0.003 | 34.930 ± 0.373 | 26.086 ± 0.304 | 0.601 ± 0.002 | 0.355 ± 0.003 |
| 563 | BG&HR | 18.711 ± 0.124 | 13.016 ± 0.097 | 0.764 ± 0.005 | 0.181 ± 0.002 | 32.747 ± 1.692 | 24.277 ± 1.739 | 0.484 ± 0.059 | 0.334 ± 0.028 |
| | BG&HRSC | 18.697 ± 0.135 | 13.001 ± 0.108 | 0.764 ± 0.005 | 0.180 ± 0.002 | 32.769 ± 1.697 | 24.285 ± 1.712 | 0.483 ± 0.062 | 0.334 ± 0.027 |
| | BG&Band | 18.683 ± 0.158 | 13.019 ± 0.107 | 0.766 ± 0.004 | 0.180 ± 0.001 | 32.618 ± 1.683 | 24.120 ± 1.612 | 0.486 ± 0.059 | 0.333 ± 0.025 |
| 570 | BG&HR | 17.381 ± 0.473 | 12.111 ± 0.322 | 0.862 ± 0.005 | 0.117 ± 0.003 | 28.701 ± 0.266 | 20.823 ± 0.212 | 0.789 ± 0.004 | 0.193 ± 0.002 |
| | BG&HRSC | 17.381 ± 0.479 | 12.113 ± 0.319 | 0.861 ± 0.004 | 0.116 ± 0.003 | 28.684 ± 0.267 | 20.816 ± 0.210 | 0.791 ± 0.004 | 0.192 ± 0.002 |
| | BG&Band | 17.395 ± 0.460 | 12.125 ± 0.312 | 0.863 ± 0.004 | 0.116 ± 0.003 | 28.736 ± 0.324 | 20.865 ± 0.247 | 0.789 ± 0.004 | 0.193 ± 0.002 |
| 575 | BG&HR | 24.030 ± 0.394 | 15.296 ± 0.251 | 0.736 ± 0.011 | 0.235 ± 0.006 | 37.351 ± 0.558 | 26.741 ± 0.517 | 0.513 ± 0.015 | 0.404 ± 0.008 |
| | BG&HRSC | 24.036 ± 0.398 | 15.306 ± 0.248 | 0.737 ± 0.011 | 0.235 ± 0.007 | 37.312 ± 0.568 | 26.695 ± 0.541 | 0.513 ± 0.014 | 0.403 ± 0.008 |
| | BG&Band | 24.027 ± 0.404 | 15.301 ± 0.253 | 0.735 ± 0.012 | 0.235 ± 0.007 | 37.303 ± 0.591 | 26.713 ± 0.557 | 0.510 ± 0.014 | 0.403 ± 0.008 |
| 588 | BG&HR | 19.126 ± 0.078 | 13.706 ± 0.077 | 0.744 ± 0.008 | 0.182 ± 0.002 | 38.258 ± 4.209 | 28.279 ± 3.472 | 0.440 ± 0.070 | 0.371 ± 0.045 |
| | BG&HRSC | 19.086 ± 0.097 | 13.738 ± 0.070 | 0.738 ± 0.010 | 0.182 ± 0.002 | 38.366 ± 4.206 | 28.427 ± 3.439 | 0.431 ± 0.069 | 0.374 ± 0.045 |
| | BG&Band | 19.314 ± 0.079 | 13.818 ± 0.086 | 0.740 ± 0.007 | 0.184 ± 0.002 | 38.500 ± 4.124 | 28.425 ± 3.387 | 0.449 ± 0.069 | 0.371 ± 0.043 |
| 591 | BG&HR | 22.332 ± 0.265 | 16.423 ± 0.284 | 0.624 ± 0.012 | 0.286 ± 0.004 | 37.346 ± 0.610 | 29.662 ± 0.566 | 0.363 ± 0.008 | 0.467 ± 0.008 |
| | BG&HRSC | 22.322 ± 0.243 | 16.420 ± 0.251 | 0.624 ± 0.011 | 0.286 ± 0.004 | 37.638 ± 0.644 | 29.936 ± 0.566 | 0.347 ± 0.014 | 0.472 ± 0.008 |
| | BG&Band | 22.528 ± 0.310 | 16.857 ± 0.342 | 0.620 ± 0.003 | 0.291 ± 0.005 | 38.161 ± 0.756 | 30.605 ± 0.595 | 0.340 ± 0.013 | 0.483 ± 0.007 |
| Avg | BG&HR | 20.178 ± 0.232 | 14.017 ± 0.181 | 0.754 ± 0.008 | 0.199 ± 0.003 | 34.796 ± 1.271 | 25.886 ± 1.117 | 0.534 ± 0.027 | 0.353 ± 0.015 |
| | BG&HRSC | 20.183 ± 0.236 | 14.035 ± 0.176 | 0.753 ± 0.008 | 0.199 ± 0.003 | 34.937 ± 1.271 | 26.030 ± 1.105 | 0.528 ± 0.029 | 0.355 ± 0.015 |
| | BG&Band | 20.259 ± 0.248 | 14.129 ± 0.193 | 0.753 ± 0.006 | 0.200 ± 0.003 | 35.041 ± 1.308 | 26.136 ± 1.117 | 0.529 ± 0.027 | 0.356 ± 0.014 |

Note. PID: patient identity; PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error; BG&HR, BG&HRSC, and BG&Band: approaches for the decision-level fusion of physical activity data with blood glucose data.




(a)




(b)

Fig. 3: Critical difference diagram showing the comparison of the no-fusion approach and feature-level physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

fusion levels of PA data with BGL data are promising in improving BGL prediction performance. Hence, it is worth noting that, by making some adjustments, similar approaches can be applied to fuse information on other affecting variables, including carbohydrates and insulin, with BGL data. More generally, apart from BGL prediction, the developed approaches can also be considered as data fusion methods for fusing two types of data for prediction tasks in related domains.

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

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

Note. PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error.

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

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

Note. PH: prediction horizon; RMSE: root mean square error; MAE: mean absolute error; MCC: Matthews correlation coefficient; SE: surveillance error.

## VI. Conclusion

The goal of this work was to contribute to finding optimal approaches for PA deployment in BGL prediction models, including the kind of PA information and the level of integration. This work comprehensively investigated leveraging PA in BGL prediction by developing different approaches for extracting various information from PA data and fusing this information with BGL data in signal, feature, and decision levels. To do so, for the signal fusion, three different combinations of automatically-recorded PA data from wristbands including (HR), (HR and SC), and (HR, SC, GSR, and ST) were fused with BGL data. Also, three feature engineering approaches, including

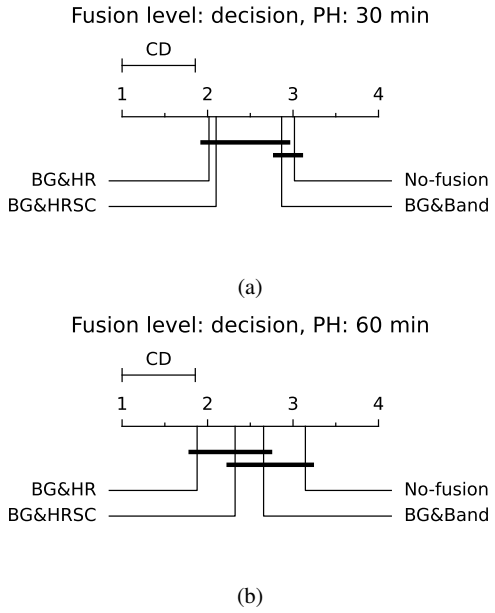



Fig. 4: Critical difference diagram showing the comparison of the no-fusion approach and decision-level physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

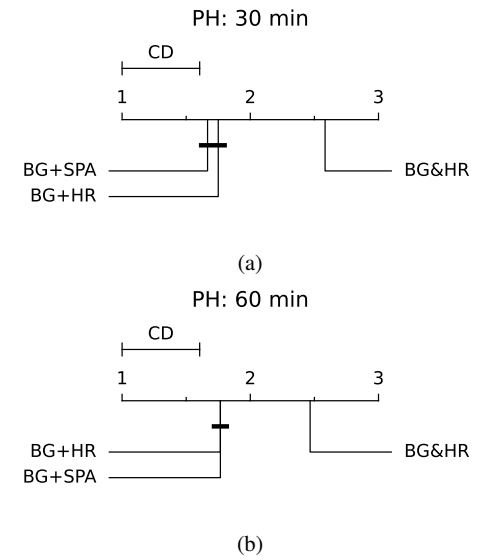



Fig. 5: Critical difference diagram showing the comparison of the effective physical activity fusion approaches against each other according to average over the distinct metrics for prediction horizon of 30 (a) and 60 (b) minutes.

subjective PA categorisation, objective PA clustering, and statistics of PA, were used to fuse feature-level PA information with BGL data. Moreover, for decision-level PA fusion, the primary decisions made by the base-learner using BGL data and PA data separately were stacked and used as inputs of the meta-learner. Three different decision-level approaches were developed based on the kind of PA data. In total, nine PA fusion approaches were developed. These approaches were compared with the no-fusion approach and also with each other. All in all, the evaluation and statistical analyses showed that fusing PA information with BGL can significantly improve the performance of BGL prediction. Among all the developed PA fusion approaches, fusing BGL with automatically recorded HR data and with categories of self-reported PA-related events outperformed the no-fusion and other PA fusion approaches. Hence, it can be concluded that the BGL prediction task can benefit from using PA bands.

Examining the developed methods for different levels of PA fusion was a preliminary investigation to find effective approaches for leveraging PA in BGL prediction. There are different ways in which this can be improved and further developed. In this work, we used different levels of PA categorised subjectively, which were useful information for BGL prediction models. Automating this procedure by automatically classifying PA data into various intensity levels would be valuable as a future direction. Also, it would be beneficial to investigate the developed approaches further by exploring more prediction models, specifically more advanced and efficient ones. Moreover, similar to the developed approaches and considering some adjustments, further investigation would involve methods for other affecting variables. It is also worth noting that testing newly developed models in as many datasets as possible can show their robustness. However, we only had access to the Ohio dataset in this work. Hence, implementing the proposed approaches on other good-quality datasets would be suggested.

## VII. Code and Data Availability

The methodologies were implemented using `Python 3.6`, `TensorFlow 1.15.0` [73], and `Keras 2.2.5` [74], deploying the following packages: `pandas` [75], `NumPy` [76], `SciPy` [77], `scikit-learn` [78], `statsmodels` [79], `scikit-posthocs` [80]. Codes implemented are accessible via the Gitlab repository. It is also possible to access the publicly available Ohio dataset through a data use agreement.

## Appendix

CD diagrams according to the evaluation metrics with a significant p-value outcome for Friedman test are displayed in Figures 6, 7, and 8. These diagrams compare the no-fusion approach with signal-level, feature-level, and decision-level PA fusion approaches. Also, CD diagrams in Figure 9 compare the effective PA fusion approaches.

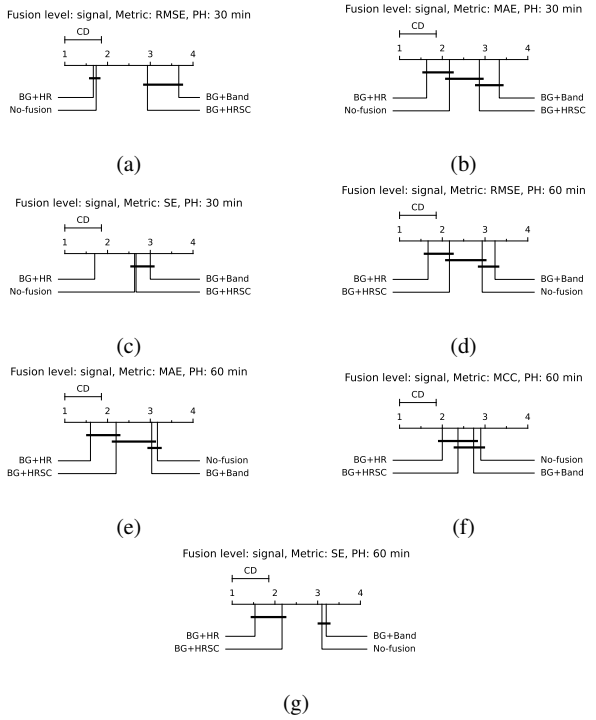



Fig. 6: Critical difference diagram showing the comparison of the no-fusion approach and signal-level physical activity fusion approaches against each other according to RMSE (a), MAE (b), and SE (c) for the prediction horizon of 30 minutes as well as RMSE (d), MAE (e), MCC (f), and SE (g) for the prediction horizon of 60 minutes.

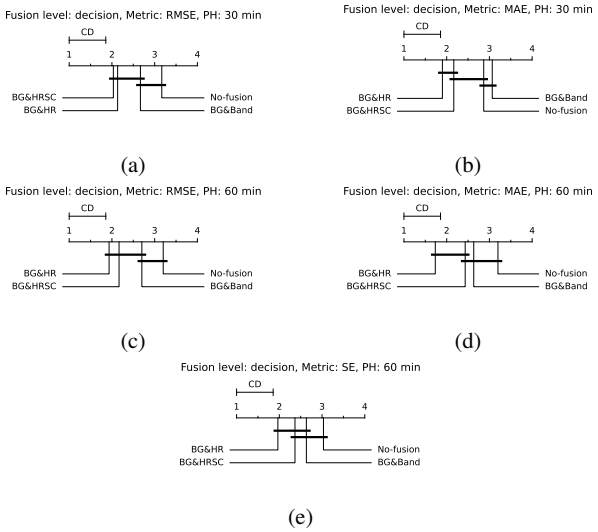



Fig. 8: Critical difference diagram showing the comparison of the no-fusion approach and decision-level physical activity fusion approaches against each other according to RMSE (a) and MAE (b) for the prediction horizon of 30 minutes as well as RMSE (c), MAE (d), and SE (e) for the prediction horizon of 60 minutes.

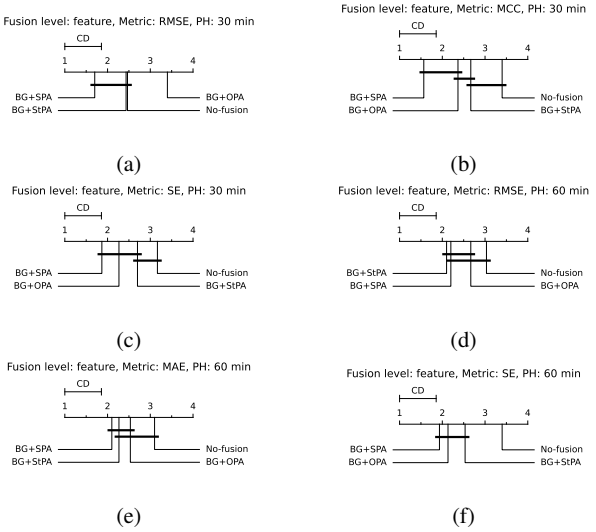



Fig. 7: Critical difference diagram showing the comparison of the no-fusion approach and feature-level physical activity fusion approaches against each other according to RMSE (a), MCC (b), and SE (c) for the prediction horizon of 30 minutes as well as RMSE (d), MAE (e), and SE (f) for the prediction horizon of 60 minutes.

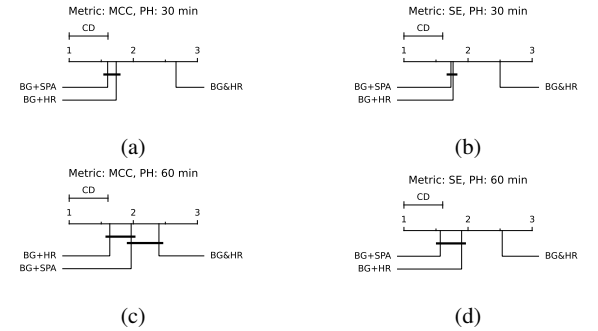



Fig. 9: Critical difference diagram showing the comparison of the effective physical activity fusion approaches against each other according to MCC (a) and SE (b) for the prediction horizon of 30 minutes as well as MCC (c) and SE (d) for the prediction horizon of 60 minutes.

## REFERENCES


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

[2] A. D. Association, "Diagnosis and classification of diabetes mellitus," *Diabetes care*, vol. 33 (Supplement 1), pp. 62–69, 2010.

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

[4] B. Balkau, L. Mhamdi, J.-M. Oppert, *et al.*, "Physical activity and insulin sensitivity: The risc study," *Diabetes*, vol. 57, no. 10, pp. 2613–2618, 2008.

[5] J. Xie and Q. Wang, "A data-driven personalized model of glucose dynamics taking account of the effects of physical activity for type 1 diabetes: An in silico study," *Journal of biomechanical engineering*, vol. 141, no. 1, p. 011 006, 2019.

[6] M. C. Riddell, I. W. Gallen, C. E. Smart, *et al.*, "Exercise management in type 1 diabetes: A consensus statement," *The lancet Diabetes & endocrinology*, vol. 5, no. 5, pp. 377–390, 2017.

[7] 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. 103 888, 2022.

[8] K. M. Metcalf, A. Singhvi, E. Tsalikian, *et al.*, "Effects of moderate-to-vigorous intensity physical activity on overnight and next-day hypoglycemia in active adolescents with type 1 diabetes," *Diabetes Care*, vol. 37, no. 5, pp. 1272–1278, 2014.

[9] J.-W. van Dijk, T. M. Eijsvogels, J. Nyakayiru, *et al.*, "Glycemic control during consecutive days with prolonged walking exercise in individuals with type 1 diabetes mellitus," *Diabetes research and clinical practice*, vol. 117, pp. 74–81, 2016.

[10] H. Tikkanen-Dolenc, J. Wadén, C. Forsblom, *et al.*, "Frequent and intensive physical activity reduces risk of cardiovascular events in type 1 diabetes," *Diabetologia*, vol. 60, no. 3, pp. 574–580, 2017.

[11] A. Z. Woldaregay, E. Årsand, S. Walderhaug, *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, 2019.

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

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

[14] H. Nemat, "Artificial intelligence in blood glucose level prediction for type 1 diabetes management," Ph.D. dissertation, University of Sheffield, 2023.

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

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

[17] T. Zhu, K. Li, J. Chen, P. Herrero, and P. Georgiou, "Dilated recurrent neural networks for glucose forecasting in type 1 diabetes," *Journal of Healthcare Informatics Research*, pp. 1–17, 2020.

[18] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2744–2757, 2020.

[19] 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. 102 923, 2021.

[20] T. Zhu, L. Kuang, K. Li, J. Zeng, P. Herrero, and P. Georgiou, "Blood glucose prediction in type 1 diabetes using deep learning on the edge," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2021, pp. 1–5.

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

[22] K. Saiti, M. Macaš, L. Lhotská, K. Štechová, and P. Pithová, "Ensemble methods in combination with compartment models for blood glucose level prediction in type 1 diabetes mellitus," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105 628, 2020.

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

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

[25] M. Wadghiri, A Idri, T. El Idrissi, and H. Hakkoum, "Ensemble blood glucose prediction in diabetes mellitus: A review," *Computers in Biology and Medicine*, vol. 147, p. 105 674, 2022.

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

[27] S. Langarica, M. Rodriguez-Fernandez, F. Nunez, and F. J. Doyle III, "A meta-learning approach to personalized blood glucose prediction in type 1 diabetes," *Control Engineering Practice*, vol. 135, p. 105 498, 2023.

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

[29] J. Daniels, P. Herrero, and P. Georgiou, "A multitask learning approach to personalised blood glucose prediction," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[30] M. De Bois, M. A. El Yacoubi, and M. Ammi, "Adversarial multi-source transfer learning in healthcare: Application to glucose prediction for diabetic people," *Computer Methods and Programs in Biomedicine*, vol. 199, p. 105 874, 2021.

[31] M. M. H. Shuvo and S. K. Islam, "Deep multitask learning by stacked long short-term memory for predicting personalized

blood glucose concentration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1612–1623, 2023.

[32] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Causality analysis in type 1 diabetes mellitus with application to blood glucose level prediction," *Computers in Biology and Medicine*, p. 106 535, 2023.

[33] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz, "Using LSTMs to learn physiological models of blood glucose behavior," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2017, pp. 2887–2891.

[34] S. Oviedo, I. Contreras, A. Bertachi, *et al.*, "Minimizing postprandial hypoglycemia in type 1 diabetes patients using multiple insulin injections and capillary blood glucose self-monitoring with machine learning techniques," *Computer methods and programs in biomedicine*, vol. 178, pp. 175–180, 2019.

[35] A. Güemes, G. Cappon, B. Hernandez, *et al.*, "Predicting quality of overnight glycaemic control in type 1 diabetes using binary classifiers," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1439–1446, 2019.

[36] H. Nemat, H. Khadem, J. Elliott, and M. Benaissa, "Data-driven blood glucose level prediction in type 1 diabetes: A comprehensive comparative analysis," *Scientific Reports*, vol. 14, no. 1, p. 21 863, 2024.

[37] M. S. M. Nordin and F. Mahmud, "Univariate and multivariate time series blood glucose prediction with lstm deep learning model," *Evolution in Electrical and Electronic Engineering*, vol. 5, no. 1, pp. 276–285, 2024.

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

[39] W. Li, Y. Guo, B. Wang, and B. Yang, "Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction," *Pattern Recognition*, vol. 136, p. 109 234, 2023.

[40] Y. Guo, D. Zhou, X. Ruan, and J. Cao, "Variational gated autoencoder-based feature extraction model for inferring disease-mirna associations based on multiview features," *Neural Networks*, vol. 165, pp. 491–505, 2023.

[41] Y. Guo, D. Zhou, P. Li, C. Li, and J. Cao, "Context-aware poly (a) signal prediction model via deep spatial–temporal neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[42] S. Yang, D. Zhou, J. Cao, and Y. Guo, "Lightingnet: An integrated learning method for low-light image enhancement," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 29–42, 2023.

[43] S. Rachakonda, S. Moorthy, A. Jain, *et al.*, "Privacy enhancing and scalable federated learning to accelerate ai implementation in cross-silo and iomt environments," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 744–755, 2022.

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

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

[46] C. Marling and R. C. Bunescu, "The OhioT1DM Dataset For Blood Glucose Level Prediction.," in *3rd International Workshop on Knowledge Discovery in Healthcare Data*, 2018, pp. 60–63.

[47] S. Cho, E. M. Aiello, B. Ozaslan, *et al.*, "Design of a real-time physical activity detection and classification framework for individuals with type 1 diabetes," *Journal of Diabetes Science and Technology*, p. 19 322 968 231 153 896, 2023.

[48] M. Cescon, D. Choudhary, J. E. Pinsker, *et al.*, "Activity detection and classification from wristband accelerometer data collected on people with type 1 diabetes in free-living conditions," *Computers in biology and medicine*, vol. 135, p. 104 633, 2021.

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

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

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

[52] B. Ozaslan, S. D. Patek, C. Fabris, and M. D. Breton, "Automatically accounting for physical activity in insulin dosing for type 1 diabetes," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105 757, 2020.

[53] N. Hobbs, S. Samadi, M. Rashid, *et al.*, "A physical activity-intensity driven glycemic model for type 1 diabetes," *Computer Methods and Programs in Biomedicine*, vol. 226, p. 107 153, 2022.

[54] A. Bertachi, C. Viñals, L. Biagi, *et al.*, "Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor," *Sensors*, vol. 20, no. 6, p. 1705, 2020.

[55] A. Beneyto, A. Bertachi, J. Bondia, and J. Vehi, "A new blood glucose control scheme for unannounced exercise in type 1 diabetic subjects," *IEEE Transactions on Control Systems Technology*, 2020.

[56] M. R. Askari, M. Ahmadasas, A. Shahidehpour, *et al.*, "Multivariable automated insulin delivery system for handling planned and spontaneous physical activities," *Journal of Diabetes Science and Technology*, vol. 17, no. 6, pp. 1456–1469, 2023.

[57] M. Sevil, M. Rashid, Z. Maloney, *et al.*, "Determining physical activity characteristics from wristband data for use in automated insulin delivery systems," *IEEE sensors journal*, vol. 20, no. 21, pp. 12 859–12 870, 2020.

[58] M. Rashid, S. Samadi, M. Sevil, *et al.*, "Simulation software for assessment of nonlinear and adaptive multivariable control algorithms: Glucose–insulin dynamics in type 1 diabetes," *Computers & Chemical Engineering*, vol. 130, p. 106 565, 2019.

[59] V. Felizardo, N. M. Garcia, N. Pombo, and I. Megdiche, "Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction–a systematic literature review," *Artificial Intelligence in Medicine*, vol. 118, p. 102 120, 2021.

[60] S. Mirshekarian, H. Shen, R. Bunescu, and C. Marling, "Lstm and neural attention models for blood glucose prediction: Comparative experiments on real and synthetic data," in *2019 41st Annual International Conference of the IEEE Engineer-*

*ing in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 706–712.

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

[62] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "Signal fragmentation based feature vector generation in a model agnostic framework with application to glucose quantification using absorption spectroscopy," *Talanta*, vol. 243, p. 123 379, 2022.

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

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

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

[66] H. Khadem, H. Nemat, M. R. Eissa, J. Elliott, and M. Benaissa, "Covid-19 mortality risk assessments for individuals with and without diabetes mellitus: Machine learning models integrated with interpretation framework," *Computers in Biology and Medicine*, vol. 144, p. 105 361, 2022.

[67] H. Khadem, M. R. Eissa, H. Nemat, O. Alrezj, and M. Benaissa, "Classification before regression for improving the accuracy of glucose quantification using absorption spectroscopy," *Talanta*, vol. 211, p. 120 740, 2020.

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

[69] H. Khadem, H. Nemat, J. Elliott, and M. Benaissa, "In vitro glucose measurement from nir and mir spectroscopy: Comprehensive benchmark of machine learning and filtering chemometrics," *Heliyon*, vol. 10, no. 10, 2024.

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

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

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

[73] M. Abadi, A. Agarwal, P. Barham, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[74] F. Chollet *et al.*, *Keras*, `https://github.com/keras-team/keras`, 2015.

[75] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 –61. DOI: `10.25080/Majora-92bf1922-00a`.

[76] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: `10.1038/s41586-020-2649-2`. [Online]. Available: `https://doi.org/10.1038/s41586-020-2649-2`.

[77] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: `10.1038/s41592-019-0686-2`.

[78] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[79] S Seabold and J Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, 2010, pp. 92–96.

[80] M. Terpilowski, "Scikit-posthocs: Pairwise multiple comparison tests in python," *The Journal of Open Source Software*, vol. 4, no. 36, p. 1169, 2019. DOI: `10.21105/joss.01169`.