DiaData: An Integrated Large Dataset for Type 1 Diabetes and Hypoglycemia Research

Beyza Cinar[1,*] and *Maria* Maleshkova[1,**]

[1]Professorship of Data Engineering, Helmut Schmidt University, Hamburg, Germany

Abstract. Type 1 diabetes (T1D) is an incurable autoimmune disorder, which needs attentive monitoring to avoid high glucose variations. Affected cannot produce sufficient insulin and depend on external insulin injections. Multiple factors impact glucose levels, which can lead to dangerous side effects of hyperglycemia (≥ 180 mg/dL) and hypoglycemia (≤ 70 mg/dL). Data analysis can significantly enhance diabetes care by discovering individual trends and enabling tailored decision support. Particularly, machine learning (ML) approaches provide early alerts and predict glucose levels. However, the main limitation in diabetes research is the unavailability of large datasets. Therefore, this study systematically integrates 15 datasets to create a comprehensive database of 2510 subjects with glucose measurements recorded every 5 minutes. In total, 149 million measurements are included (Euglycemia (58.3%), Hyperglycemia (37.5%), and Hypoglycemia (4.2%)). Moreover, two sub-databases are extracted, including demographics or heart-rate data. The integrated dataset provides an equal distribution of sex and a variety of age levels. As a further contribution, data quality is assessed, revealing that missing values and data imbalance present a significant challenge. Thus, the application of ML models necessitates appropriate preprocessing methods.

1 Introduction

Type 1 diabetes (T1D) is one of the most prevalent chronic diseases among children and adolescents, presenting substantial challenges in healthcare management. IDF Diabetes Atlas estimates that over 9 million people worldwide were living with T1D in 2024. Of these, approximately 1.8 million were under the age of 20, while 1.1 million were aged 60 years or older [1, p.72]. T1D is an autoimmune disease characterized by insulin deficiency, which is why affected depend on lifelong insulin administration. Since glucose levels cannot be regularized normally, patients with T1D usually have unsteady and increased glucose levels, which is referred to as hyperglycemia (≥180 mg/dL). Hyperglycemia can cause diabetic ketoacidosis, and microvascular and macrovascular comorbidities [2, 3]. Thus, therapy includes tailored insulin injections or the use of insulin pumps, which, however, have a side effect of hypoglycemia (≤70 mg/dL). Moreover, lifestyle is restricted since meal intake and activity also impact hypoglycemia [2, 4]. Nowadays, biomedical sensors such as continuous glucose measurement (CGM) devices enhance diabetes therapy. These devices measure

*e-mail: cinarb@hsu-hh.de
**e-mail: maleshkm@hsu-hh.de

glucose levels in the interstitial fluid through a microneedle placed under the skin [3, 5]. Artificial intelligence (AI) methods like machine learning (ML) and deep learning (DL) can be trained on CGM data to forecast glucose values or predict the risk of adverse events. Notably, neural networks (NNs) provide stable results as they can identify patterns automatically from time series data [5, 6]. Hence, data analysis can significantly improve tailored diabetes management by discovering individual trends and risks. However, these models need to be trained on large (multivariate) datasets. Larger datasets can further support the development of evidence-based clinical guidelines, enable risk stratification, and enhance the understanding of disease patterns and demographic variations. The limited availability of such datasets remains the leading research gap and constrains the potential of AI-driven approaches. Typically, models are trained on only 5 to 20 subjects [6], which deliver data volumes that are insufficient for most ML approaches but are also under-representative for the patient groups.

This study addresses this gap with the objective to enhance diabetes care and to enable the development of robust algorithms that mitigate bias. We present DiaData, a large CGM dataset, by systematically integrating 15 datasets to enable effective training on ML and DL models. In particular, this study makes the following contributions: 1) It provides DiaData, a large dataset with glucose measurements recorded every 5 minutes (15 datasets). 2) It provides a sub-dataset including demographics of age and sex (10 datasets). 3) It provides a sub-dataset including heart-rate data (3 datasets). 4) Finally, it investigates the quality of the dataset and explores cohort statistics.

The remainder of this work is as follows. Section 2 highlights related work and marks research gaps. Section 3 presents the methods used for integrating the dataset. Section 4 assesses data quality and investigates bias in demographics. Findings are discussed in section 5, and lastly, section 6 provides a conclusion.

2 Related Work

In recent years, AI-based approaches have been increasingly explored to enhance diabetes management and prevent serious complications. These encompass glucose forecasting, hypoglycemia prediction, and insulin optimization [6]. Reviews show that various algorithms are approached, with neural network and deep learning-based models reported to be superior for prediction. However, a major limitation is the limited access to public datasets, which undermines the full exploration of AI's potential [3, 6]. Without large, representative databases, neither individual nor population-based models can be effectively trained. Felizardo et al. further note that the majority of studies train on datasets comprising only 5 to 20 subjects, with the OhioT1DM dataset (12 subjects) being the most popular [6]. Other usually utilized public datasets for T1D include the D1NAMO (9 subjects) [7], ShanghaiT1D (25 subjects) [8], and HUPA-UCM datasets (18 subjects) [9]. Cinar et al. demonstrate that DL models trained on small datasets tend to overfit in hypoglycemia classification. While the OhioT1DM dataset enables short-term prediction, long-term forecasting remains challenging [10]. Similarly, smaller datasets like D1NAMO do not support generalizable or validated approaches [11].

In the context of data integration methods, Dong et al. propose a three-step approach for particularly big data. Schema mapping identifies correspondences between global and local schemata. Record linkage merges structured records with the same schema. Finally, data fusion integrates all data [12]. Sourlos et al., focusing on health data, give more detailed instructions. First, the specific use case and clinical target must be defined to select the most relevant features. The integrated dataset should contain ground truth and may provide labels. Since data can be collected in different settings, data harmonization is essential. Also, initial analyses of data and subgroups should address fairness and potential bias. Finally, baseline performance can be evaluated by comparing models on the benchmark dataset [13].

Notably, two fused datasets collecting glucose were presented for diabetes research. First, Arévalo et al. curated a dataset from the MIMIC III database. They matched glucose and insulin values collected in the intensive care unit from electronic health records. The motivation was to analyze insights into glycemic range targets impacting the survival rate for critically ill patients [14]. However, no CGM device was used. Second is GlucoBench, which integrates datasets including CGM data. Sergazinov et al. fused 5 different databases comprising type 1, type 2 (T2), or a mixture of T2 and no diabetes subjects with the aim to enhance prediction accuracy. Datasets were included if they contained at least five subjects, each with non-missing CGM measurements not exceeding 16 consecutive hours. Subjects with poor data quality were removed as well. Moreover, only measurements in the relevant range of 20 mg/dL to 400 mg/dL were included. One disadvantage is that the single datasets have different sizes and demographics, which can lead to bias in the integrated dataset. They further state that different diabetes types may impact prediction accuracy. Finally, data quality is improved by imputing missing values with linear interpolation and standardizing data using min-max-scaling [3]. While CGM data are included, GlucoBench may not be effective for hypoglycemia prediction due to the limited scale of its T1D subgroup.

3 Methodology

This study systematically integrates various datasets collecting CGM data of patients with T1D. The objective is to enhance glucose management by improving the performance of prediction models. An integrated dataset can be used for the training of generalized models or the validation of models for glucose forecasting. To enable generalization, we aimed to include at least 1000 different subjects and represent a variety of age levels.

Figure 1 presents the applied systematic process of dataset integration. Initially, data were collected and individually sampled at a uniform frequency of 5 minutes. The datasets were then integrated into one main dataset comprising all subjects' IDs, timestamps, and the corresponding glucose values. Then, we extracted two sub-databases from the main database: 1) sub-database I included data on age and sex, and 2) sub-database II included heart-rate values. This integration method produces three complete datasets with fewer missing values, rather than relying on a single, comprehensive dataset that suffers from data incompleteness.

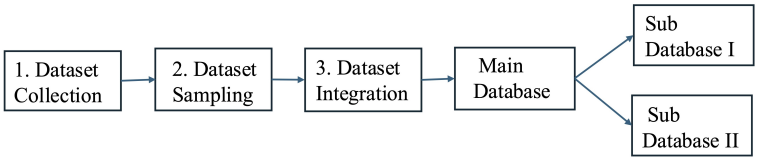


Figure 1: Applied Process of Dataset Integration

1. Dataset Collection: As illustrated in Figure 2, which presents a common data integration framework [12, 13], the initial step was to identify the use cases for the database, which are hypoglycemia classification and glucose forecasting. Then, we collected datasets meeting the following conditions: 1) They should include CGM data. 2) They should include only patients with T1D. 3) They should be mainly collected under free-living conditions.

The final cohort consisted of 15 datasets and 2510 patients, yielding over 800,000 days of data in total. Individual statistics of the single datasets can be seen in Table 1. Analysis of the minimum, maximum, and mean number of collected values indicates variability in data availability, with some subjects having limited records. The number of patients per dataset ranges from a minimum of 1 to a maximum of 736. Notably, T1DiabetesGranada (T1DGranada) and

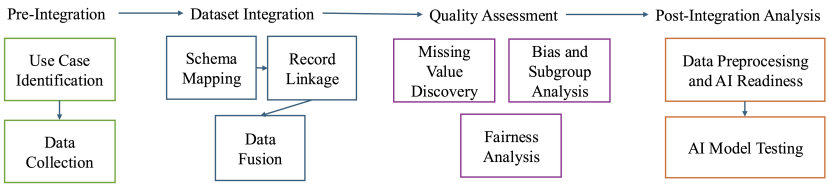


Figure 2: Data Integration Framework

Table 1: Overview of the Single Datasets

Data	**Subjects**	**Mean Values**	**Min Days**	**Max Days**	**Mean Days**	**Demo-graphics**	**HR**
T1DGranada [15]	736	87419	6	1531	443	✓	
DiaTrend [16]	54	134833	30	2179	564	✓	
CITY [17]	153	54004	10	709	348		
DLCP3 [17]	169	44701	11	867	179	✓	
HUPA-UCM [18]	24	21265	7	1152	174		✓
PEDAP [17]	103	70389	13	480	286	✓	
RBG [17]	226	65047	72	820	277		
RT-CGM [17]	451	42809	6	451	355	✓	
SENCE [17]	144	57687	11	419	342	✓	
SHD [17]	200	3220	5	58	14	✓	
WISDM [17]	206	51307	12	496	356	✓	
ShanghaiT1D [19]	16	2940	3	13	9	✓	
D1NAMO [20]	9	884	0	4	3		✓
DDATSHR [21]	18	16011	15	87	72	✓	✓
T1GDUJA [22]	1	41292	246	246	246		

RT-CGM datasets include more subjects than the average, whereas T1GDUJA, DDATSHR (dataset diabetes adolescents time-series with heart-rate), D1NAMO, and ShanghaiT1D contain significantly fewer. Consequently, variations in the average number of glucose values, monitoring duration, sampling frequency, and the volume of usable and hypoglycemic data across datasets and subjects may introduce a risk of bias.

2. Dataset Sampling: After the corpus generation, we defined global and local schemata, which can be seen in Figure 3. To enable seamless integration, local column names were renamed. For each relation $X \in \{X_1, X_2, \ldots, X_n\}$, we applied $\rho_{X(a',b',c',d',e')}(\pi_{a,b,c,d,e}(X))$ in which X is the database, (a', b', c', d', e') are (GlucoseCGM, PtID, Sex, Age, HR), and (a, b, c, d, e) are the individual old values. The single databases were preprocessed separately.

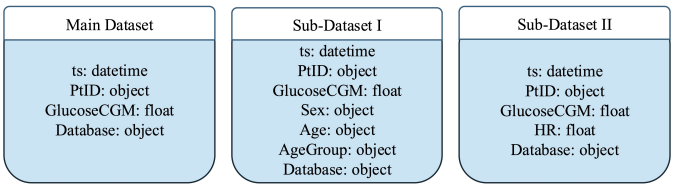


Figure 3: Database Schema

First, misaligned rows of one subject from the HUPA-UCM database were manually corrected. To the first three lines in the file, we added four extra semicolons at the end of the last column. Second, to enable resampling, we created a timestamp column named "ts" for each

database. For datasets reporting the date and the time as separate columns, the attributes were concatenated and converted to a datetime object. The ReplaceBG (RBG) dataset and Severe Hypoglycemia Dataset (SHD) only reported the number of days passed since admission and the measured time as separate columns. Thus, we created a valid datetime column by adding the reported days to a randomly generated date (01.01.2024).

The timestamps were rounded to 5 minutes to have uniform timestamps across each subject and database, and resulting duplicates were removed. Then, we resampled the timestamps to 5-minute intervals. For databases using multiple sensors for the study duration (DDATSHR and RT-CGM), the measurement frequency was identified before resampling. Dataframes including heart-rate (HR) were only rounded to 1-minute intervals and duplicates were removed since values were typically reported each second. Thereafter, they were aligned to the glucose measurements with an outer join. In total, we undersampled 5 datasets, which were collected with frequencies of 10 or 15 minutes. Undersampling is appropriate in this context, as 15-minute gaps can be effectively addressed via imputation or interpolation [23]. We imputed produced missing values between collected parameters with feedforward filling. In contrast, standardizing all data to 15-minute intervals would reduce volume and risk losing representativeness. Moreover, distinguishing between 5, 10, and 15 minutes is crucial for short-term glucose forecasting and hypoglycemia classification.

3. Dataset Integration: Two datasets reported glucose in mmol/L (D1NAMO and DDATSHR), thus we converted the glucose values to mg/dL by multiplying with 18.02. In the CITY and D1NAMO datasets, the glucose columns contained CGM-derived and manually recorded glucose values. These were separated into distinct columns, "GlucoseCGM" and "GlucoseManual". For the purposes of this study, we retained only "GlucoseCGM" and excluded "GlucoseManual". Manual glucose values may prove valuable in future work for imputation, calibration, or validation of CGM accuracy. The DDATSHR and HUPA-UCM datasets furthermore store historic and scan glucose values separately. For the integration into a single column before resampling, we substituted the historical values with their corresponding, time-aligned scan glucose readings.

Then, we extracted the age and sex from the corresponding dataframes. For the T1DGranada dataset, we computed the age from the reported birthdate of the patient and the timestamps. The SHD dataset mentioned that patients at least aged 60 were chosen for the study. Thus, we stored the age as a string of "60-100", assuming 100 to be the maximum age. All extracted values were aligned to the timestamps and patient IDs. For all databases, we stored sex as "F" for females and "M" for males. Finally, we added a "Database" column to each database to enable the re-identification of the source of data.

Once semantic equality was established among the datasets, we aligned them, adhered to the predetermined global schema, and then fused the data. While integrating the datasets, a column for "AgeGroups" was created since some ages were reported as strings in ranges. We defined bins of 0−2, 3−6, 7−10, 11−13, 14−17, 18−25, 26−35, 36−55, 56+ based on the reported age ranges. Finally, we removed patients having only null values for the "GlucoseCGM" column. Sourlos et al. further suggest providing labels to aim for a fair and AI-ready dataset [13]. However, the glucose values already serve as ground truth, and depending on the ML method and algorithm, different data structures or labels may be required.

Finally, Table 2 summarizes the fused datasets. The main dataset (15 datasets) includes 149 million usable, non-missing CGM data points, with an average of one year of data per individual. Sub-dataset I (10 datasets) shows similar characteristics, whereas sub-dataset II (3 datasets) comprises only 0.5% of the main dataset, with an average of 108 days of glucose values per subject. Therefore, even if training models with multiple wearable data increases performance [6], sub-dataset II may perform worse. This distribution underlines the need for more multivariate datasets, including vital parameters.

Table 2: Summary of the Integrated Datasets

Dataset	Subjects	Total Values	Values ≤ 70	Mean days	Mean values
Main Dataset	2510	149112702	6195639	328	59407
Sub-Dataset I	2096	125585996	5327243	335	59916
Sub-Dataset II	51	806547	41284	108	15814

4 Quality Assessment

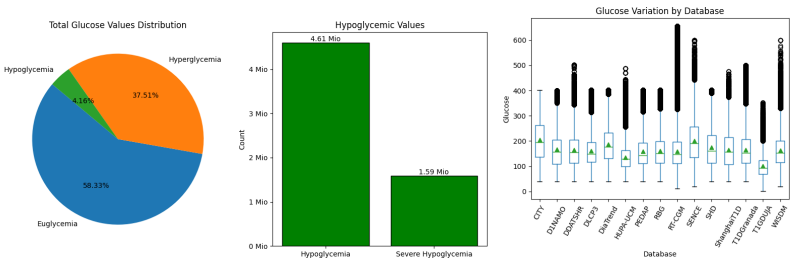
This section presents the quality assessment of the data integration framework illustrated in Figure 2. In particular, we analyze the glucose characteristics of the dataset and assess potential sources of bias. As previously indicated in Table 1, the over-representation of certain subjects suggests a potential risk of bias.

4.1 Glucose Characteristics

The main dataset comprises 6.2 million hypoglycemic, 55.9 million hyperglycemic, and 86.9 million target range (euglycemic) data points. This introduces a significant imbalance, particularly for hypoglycemia, which makes up only 4.16% of all collected glucose values, as visualized in Figure 4a. Moreover, the distribution between level 1 (≤70 mg/dL) and level 2 (≤54 mg/dL) hypoglycemia is imbalanced. Thus, using all values could lead to a bias of increased glucose values, while the prediction of hypoglycemia could be more challenging without additional preprocessing like undersampling or oversampling using synthetic data. In the context of hypoglycemia classification restricted to hypoglycemic data points, a total of 21 thousand days are available from 2501 subjects, excluding missing values.

Conversely, hyperglycemic values are well represented with 37.51%, underscoring the difficulty of maintaining glucose levels within the target range in affected subjects. Prolonged and frequent hyperglycemia can lead to vascular comorbidities and should be minimized [2].

Figure 4b illustrates the boxplots of glucose ranges across the single datasets. In particular, a lot of outliers are detected, indicating artifacts in CGM device measurements between 400-700 mg/dL. Therefore, the confidence rate of the CGM devices should be investigated to filter outliers. Similarly, very low glucose values of zeroes should be filtered because these are physiologically implausible, as observed in the RT-CGM and T1GDUJA datasets. Notably elevated glucose levels are observed in the RT-CGM, SENCE, and WISDM datasets, while the CITY, SHD, and DiaTrend datasets exhibit the fewest outliers. In most datasets, the third quartile predominantly comprises hyperglycemic data points, and the whiskers to the maximum values are significantly longer than to the minimum, underscoring the prior findings. It can be seen that the median is not always centered and that the data are skewed toward hyperglycemia. If not centered, the median is closer to the lower values. In addition, across the datasets, a difference in the length of boxes indicates variation between the single datasets, which could depend on age or the study settings. In particular, the T1GDUJA and HUPA-UCM datasets show less variance, demonstrating that the subjects experience fewer glucose ranges. While values in the HUPA-UCM dataset are typically in the euglycemic range, they behave in the lower values in the T1GDUJA dataset. In contrast, the CITY, SENCE, and SHD datasets present more variance, indicating larger glucose ranges, of which most behave in hyperglycemic ranges. Still, the majority of datasets are mainly in the euglycemic range. These observations conclude that DiaData encompasses a diverse range of subject characteristics, which could improve its generalization performance. However, proper preprocessing is essential before using the data for training.

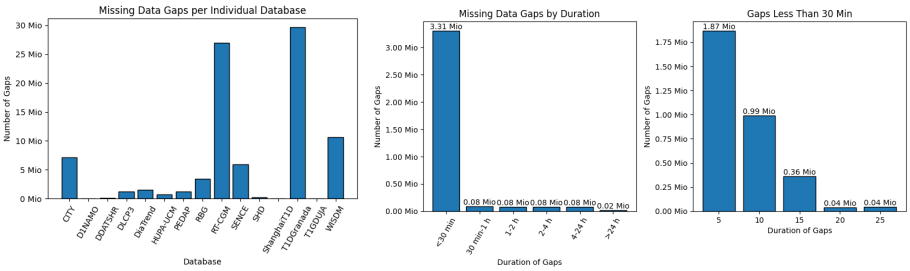


(a) Distribution of Glucose Values (b) Variance of Glucose Levels

Figure 4: Analysis of Glucose Levels in the Main Dataset

4.2 Missing Values

Missing values can occur due to sensor errors, behavioral factors, and not wearing the wearable device continuously. Examining the quality of the dataset from Figure 5a, it can be seen that missing values are a substantial problem of the database. Most studies probably have not measured data continuously during the whole study duration across all subjects. Notably, up to 88.8 million glucose values are missing, leading to a gap of approximately 308 thousand days. As visualized in Figure 5a, the datasets holding most of the subjects have the majority of missing values (T1DGranada and RT-CGM). Examining the missingness of values in more detail, Figure 5b demonstrates that the majority of gaps occur between 10 and 30 minutes, indicating sensor or environmental errors. Most of these are 5-minute gaps and can be effectively imputed using appropriate methods such as linear interpolation, which performs well for short-term gaps [23]. Consequently, despite substantial missing data, proper preprocessing can improve data quality. In contrast, gaps exceeding 30 minutes are more evenly distributed, implying that the sensor was either improperly attached or not worn. Larger gaps of 4 to more than 24 h indicate prolonged sensor removal and should probably be removed.

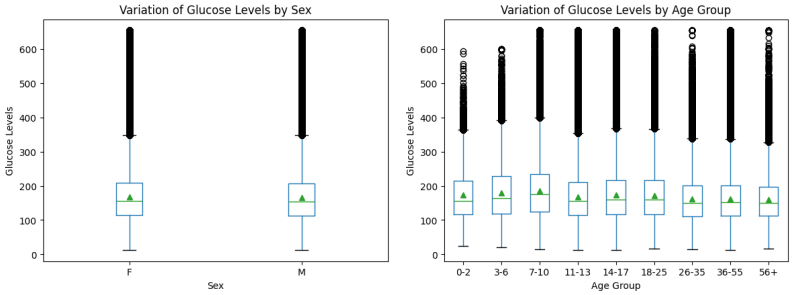


(a) Missing Value Count per Dataset (b) Missing Value Count of the Main Dataset

Figure 5: Analysis of Missing Values

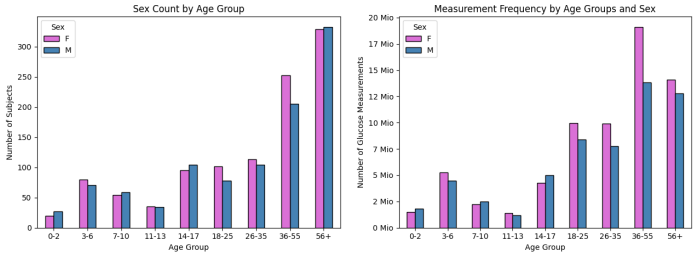
4.3 Demographics

Turning now to the analysis of sub-dataset I, sex is equally distributed, with females comprising 51.6% (66 more females out of 2096 subjects) and males 48.4%. Likewise, the total number of measurements is similarly balanced, with females contributing 53.95% and males 46.05%. Females have up to 9.9 million more data points in total, which could potentially



(a) Variance of Glucose Across Sex (b) Variance of Glucose Across Age

Figure 6: Boxplot Representation of Glucose Levels Across Demographic Groups



(a) Sex Ratio by Age Group (b) Measurement Frequency

Figure 7: Distribution of Demography

introduce a bias. Nevertheless, Figure 6a reveals that the boxplots of both sexes are very similar, indicating that the glucose ranges do not differ much, with similar medians and only a slightly higher mean in males.

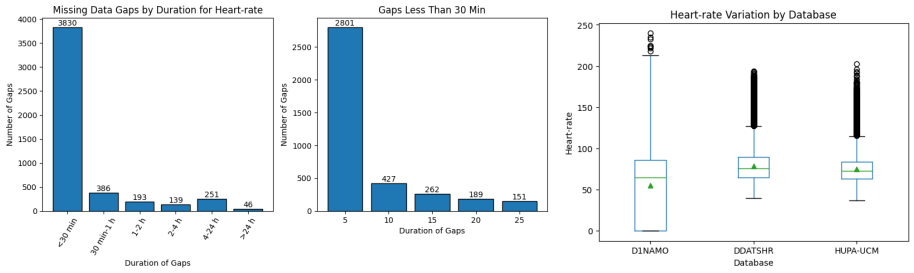
Figure 7 presents the age group distribution, with Subfigure 7a highlighting that adults aged 18 to 55 are the most represented group, comprising up to 855 subjects. More females are included in these age groups, especially between 36 and 55 years. The second largest group comprises elderly adults (≥56 years) with up to 662 subjects, displaying an equal sex distribution. The least present are newborns, children, and teenagers, with 579 subjects, also with a balanced sex ratio. Data from children and teenagers were collected from 6 studies of the integrated dataset. The uneven subject distribution across age groups may introduce bias when analyzing the full cohort, whereas focusing on individual age groups may not include enough data to be fully representative. Subfigure 7b also reveals demographic differences between the ratio of glucose measurements, while most present are females between 36 and 55 years. Other age groups show a similar distribution of values among the sexes. Even if the group aged more than 56 has more subjects, the measurement frequency is similar for males but significantly less for females of the 36 and 55 groups. Young children (3–6 years), young adults (18–25), and adults (26–35) have very similar distributions of glucose measurements, and also among sexes. The least representative are newborns (0-2) and children (11-13).

Finally, examining bias possibility in more detail, Figure 6b presents boxplots illustrating the skewness of measured glucose values across different age groups. While the variance, mean, minimum, and maximum values are very similar for adults (≥26 and ≥56 years), discrepancies can be observed between newborns, children, teenagers, and adults. Therefore, based on the glucose characteristics, the subjects can be grouped into 0-10, 11-25, and 26 to more than 56 years. The first group shows increased values with increased maximums, the second group has slightly decreased values, and the last group tends to have more controlled

glucose ranges with less variance and smaller maximums. For the earlier ages until 10 years, more variance can be observed than for the adults and elderly, while the mean is nearer to the lower values. Conclusively, age groups can indeed propose a bias since glucose values show different behaviors and patterns in each group. It is proposed to train a generalized model but also to compare among children, teenagers and young adults, and adults and the elderly.

4.4 Heart Rate Characteristics

Lastly, we examine missing values and the distribution of heart-rate data in sub-dataset II. Likewise, most missing values are short-term gaps under 30 minutes, with the majority being 5-minute gaps, as shown in Figure 8a. The boxplots in Figure 8b indicate lower variance and fewer outliers compared to glucose measurements. The DDATSHR and HUPA-UCM datasets have similar variance, means, and minimum values, while the DDATSHR dataset shows higher variance and a greater maximum. Most values are between 80 and 90 beats per minute (bpm), with outliers ranging from 120 to 200 bpm. However, the D1NAMO dataset displays greater variance, likely due to differences in sensor brands. A lot of sensor errors are recorded as zero values rather than introduced as missing values. These observations could irritate the boxplot, skewing it to lower bounds. In addition, outliers exceed 210 bpm in the D1NAMO dataset, marking a great difference from the other datasets. Thus, heart-rate values could need further preprocessing, such as normalization and standardization, to enable the use of all datasets.



(a) Missing Value Count of Sub-Dataset II (b) Variance of Heart-Rate Data

Figure 8: Analysis of Heart-Rate Values

5 Discussion

This study provides DiaData, an integrated dataset of 15 datasets collected from patients with T1D involving 149 million CGM values, demographics, and heart-rate. The primary advantage of DiaData over existing datasets lies in its heterogeneous cohort and large volume of CGM data, which enables the development of both generalized population models and individualized predictive frameworks. DiaData is well-suited for training advanced DL models, such as LSTM networks for glucose forecasting, which are prone to overfitting on smaller datasets. Its large scale enables data-driven models to capture a broader range of patterns and enables model fine-tuning for individualized predictions. In particular, long-term data from some subjects could rather be used for model individualization while retraining generalized models with transfer learning. Including such individuals in population-based models may bias overall model performance. Personalization can enhance model performance and deliver tailored decision support. This is particularly beneficial for small datasets limited to hypoglycemic values, in which transfer learning can improve short-term predictions [10].

Notably, DiaData includes 6 million hypoglycemic data points, making it a valuable resource for hypoglycemia research. However, the imbalance among glucose states poses a challenge for data-driven applications since hypoglycemia constitutes only 4% of the dataset. This imbalance can hinder the accurate prediction and classification of hypoglycemia because the model may be biased toward the majority classes (euglycemia and hyperglycemia). To address this limitation, balancing techniques like oversampling, undersampling, or synthetic data augmentation should be considered. Additionally, training expert models on specific glucose ranges may improve predictive accuracy for underrepresented states. Furthermore, heart-rate data are less present, stressing the need for more multivariate datasets.

DiaData includes a balanced representation of both sexes and a wide range of age groups, from newborns to older adults, while adults over the age of 36 are over-represented. Moreover, notable variations can be identified across age groups, with each group demonstrating different variances in glucose levels. These could introduce a bias impacting the performance of ML models. The observed bias may also be due to the defined age groups, with younger cohorts divided into narrower intervals of 2-3 years, whereas older cohorts were grouped into broader age ranges. This uneven stratification may have introduced structural inequalities in the demographic distribution. However, the narrower grouping of younger cohorts is justified by their fast biological and physiological development, which requires a narrower differentiation. Based on the glucose characteristics and the boxplot analysis, a further classification into children, teenagers+young adults, and adults+elderly adults is suggested.

Integrating all datasets has shown that different sensors and estimation methods make seamless integration of data challenging. Currently, there is no common standard, and data are not in an AI-ready format. Continuous sensor data cause a lot of missing values, which require proper imputation methods. Notably, most missing values involve 5-minute gaps, which could have occurred due to sensor and environmental errors. Moreover, different gap sizes of consecutive missing values could require different imputation methods as stated in [23], since glucose behavior is not linear. It is proposed to remove gaps that are larger than 4 hours and impute the remaining missing dataset.

Improving DiaData's quality and providing an AI-ready format serve as future aims. Additionally, future work will benchmark DiaData's utility by comparing model performance against state-of-the-art methods for hypoglycemia prediction, evaluating the impact of its larger data volume and diverse population. Model testing was beyond the scope of this study, as we demonstrated that the dataset requires appropriate preprocessing before evaluation.

6 Conclusion

This study integrated 15 databases collecting CGM data into DiaData, a large glucose database, which can be used to train ML or DL models to forecast glucose or to predict adverse events. Particularly, we presented a main database comprising 2510 patients. Two additional sub-databases can be extracted from the main database, which either include demographics (10 databases with 2096 subjects) or heart-rate measurements aligned with the glucose estimations (3 databases with 51 subjects). Characteristics of the curated databases show that: 1) For the main database, the distribution among hyperglycemia, euglycemia, and hypoglycemia is imbalanced, with the latter being under-represented. Most missing values in each database are of short-term gaps of less than 30 minutes, indicating sensor errors, misplacement, or sensor changes. Those can be imputed with preprocessing and imputation methods. 2) The distribution among sexes is equal in the integrated dataset, while the distribution across age groups differs. 3) Only 3 databases include heart-rate values, which could lead to decreased performance. Future work will focus on enhancing data quality by imputing missing values, removing large data gaps, normalizing, and correcting sensor errors.

Data and Code Availability

The datasets used in this study were obtained from multiple third-party sources. Due to licensing restrictions, we do not redistribute the full integrated dataset. However, all preprocessing and integration scripts are publicly available at https://github.com/Beyza-Cinar/DiaData, along with instructions for obtaining the original datasets. A subset of the dataset can be downloaded from the following source: [24].

The sources of subsets of the data are the Barbara Davis Center, Jaeb Center for Health Research, Joslin Diabetes Center, T1D Exchange, University of Colorado, and the University of Virginia. Retrieved from: https://public.jaeb.org/dataset. The analyses, content, and conclusions presented herein are solely the responsibility of the authors and have not been reviewed or approved by the before mentioned institutions.

References

[1] International Diabetes Federation, IDF Diabetes Atlas, 11th edn. (International Diabetes Federation, Brussels, Belgium, 2025), ISBN 978-2-930229-96-6, https://diabetesatlas.org/resources/idf-diabetes-atlas-2025/

[2] N. Nilam, S. M., P. N., Therapeutic Modelling of Type 1 Diabetes (InTech, 2011), ISBN 9789533077567, http://dx.doi.org/10.5772/21919

[3] R. Sergazinov, E. Chun, V. Rogovchenko, N. Fernandes, N. Kasman, I. Gaynanova, GlucoBench: Curated List of Continuous Glucose Monitoring Datasets with Prediction Benchmarks (2024), version Number: 1, https://arxiv.org/abs/2410.05780

[4] A. Piersanti, B. Salvatori, C. Göbl, L. Burattini, A. Tura, M. Morettini, A Machine-Learning Framework based on Continuous Glucose Monitoring to Prevent the Occurrence of Exercise-Induced Hypoglycemia in Children with Type 1 Diabetes, in *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)* (IEEE, 2023), p. 281–286, http://dx.doi.org/10.1109/CBMS58004.2023.00231

[5] M. Vettoretti, G. Cappon, A. Facchinetti, G. Sparacino, Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors, Sensors **20**, 3870 (2020). 10.3390/s20143870

[6] V. Felizardo, N.M. Garcia, N. Pombo, 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 **118**, 102120 (2021). 10.1016/j.artmed.2021.102120

[7] F. Dubosson, J.E. Ranvier, S. Bromuri, J.P. Calbimonte, J. Ruiz, M. Schumacher, The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management, Informatics in Medicine Unlocked **13**, 92 (2018). 10.1016/j.imu.2018.09.003

[8] Q. Zhao, J. Zhu, X. Shen, C. Lin, Y. Zhang, Y. Liang, B. Cao, J. Li, X. Liu, W. Rao et al., Chinese diabetes datasets for data-driven machine learning, Scientific Data **10**, 35 (2023). 10.1038/s41597-023-01940-7

[9] J.I. Hidalgo, J. Alvarado, M. Botella, A. Aramendi, J.M. Velasco, O. Garnica, HUPA-UCM diabetes dataset, Data in Brief **55**, 110559 (2024). 10.1016/j.dib.2024.110559

[10] B. Cinar, J.D. Onwuchekwa, M. Maleshkova, Deep learning-based hypoglycemia classification across multiple prediction horizons (2025), https://arxiv.org/abs/2504.00009

[11] B. Cinar, F. Grensing, L. Van Den Boom, M. Maleshkova, in *Artificial Intelligence in Healthcare*, edited by X. Xie, I. Styles, G. Powathil, M. Ceccarelli (Springer Nature Switzerland, Cham, 2024), Vol. 14975, pp. 98–109, ISBN 978-3-031-67277-4

978-3-031-67278-1, series Title: Lecture Notes in Computer Science, https://link.springer.com/10.1007/978-3-031-67278-1_8

[12] X.L. Dong, D. Srivastava, Big data integration, in *2013 IEEE 29th International Conference on Data Engineering (ICDE)* (IEEE, Brisbane, QLD, 2013), pp. 1245–1248, ISBN 978-1-4673-4910-9 978-1-4673-4909-3 978-1-4673-4908-6, http://ieeexplore.ieee.org/document/6544914/

[13] N. Sourlos, R. Vliegenthart, J. Santinha, M.E. Klontzas, R. Cuocolo, M. Huisman, P. Van Ooijen, Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology, Insights into Imaging **15**, 248 (2024). 10.1186/s13244-024-01833-2

[14] A. Robles Arévalo, J.H. Maley, L. Baker, S.M. da Silva Vieira, J.M. da Costa Sousa, S. Finkelstein, R. Mateo-Collado, J.D. Raffa, L.A. Celi, F. DeMichele, Data-driven curation process for describing the blood glucose management in the intensive care unit, Scientific Data **8**, 80 (2021). 10.1038/s41597-021-00864-4

[15] C. Rodriguez-Leon, M.D. Aviles-Perez, O. Banos, M. Quesada-Charneco, P.J. Lopez-Ibarra Lozano, C. Villalonga, M. Munoz-Torres, T1diabetesgranada: a longitudinal multi-modal dataset of type 1 diabetes mellitus, Scientific Data **10** (2023). 10.1038/s41597-023-02737-4

[16] T. Prioleau, A. Bartolome, R. Comi, C. Stanger, Diatrend: A dataset from advanced diabetes technology to enable development of novel analytic solutions, Scientific Data **10** (2023). 10.1038/s41597-023-02469-5

[17] Jaeb Center for Health Research, Diabetes datasets - public data archive, https://public.jaeb.org/datasets/diabetes (2025), accessed: 2025-04-23

[18] J. Alvarado, Hupa-ucm diabetes dataset (2024), https://data.mendeley.com/datasets/3hbcscwz44/1

[19] J. Zhu, Diabetes datasets-shanghait1dm and shanghait2dm (2022), https://figshare.com/articles/dataset/Diabetes_Datasets-ShanghaiT1DM_and_ShanghaiT2DM/20444397/3

[20] F. Dubosson, Jean-Eudes Ranvier, S. Bromuri, J.P. Calbimonte, J. Ruiz, M. Schumacher, The open d1namo dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management (2018), https://zenodo.org/record/5651217

[21] ICT Innovaties Zorg, Dataset - diabetes adolescents time series with heart rate, https://github.com/ictinnovaties-zorg/dataset-diabetes-adolescents-time-series-with-heart-rate/tree/main (2025), accessed: 2025-04-23

[22] J.F. Gaitán Guerrero, J.L. López Ruiz, C. Martínez Cruz, M. Espinilla Estévez, T1gduja: Glucose dataset of a patient with type 1 diabetes mellitus (2024), https://zenodo.org/doi/10.5281/zenodo.11284018

[23] S. Lin, X. Wu, G. Martinez, N.V. Chawla, Filling Missing Values on Wearable-Sensory Time Series Data (Society for Industrial and Applied Mathematics, 2020), p. 46–54, ISBN 9781611976236, http://dx.doi.org/10.1137/1.9781611976236.6

[24] B. Cinar, M. Maleshkova, Diadata: an integrated large dataset for type 1 diabetes and hypoglycemia research (2025). 10.24405/20048