mathematics

MDPI

Article

The Impact of Missing Continuous Blood Glucose Samples on Machine Learning Models for Predicting Postprandial Hypoglycemia: An Experimental Analysis

Najib Ur Rehman [1], Ivan Contreras [1,2,*], Aleix Beneyto [1] and Josep Vehi [1,3,*]

[1] Modeling & Intelligent Control Engineering Laboratory, Institute of Informatics and Applications, Universitat de Girona, 17003 Girona, Spain; syed.najib@udg.edu (N.U.R.); aleix.beneyto@udg.edu (A.B.)
[2] Professor Serra Húnter, Universitat de Girona, 17003 Girona, Spain
[3] Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), 17003 Girona, Spain
* Correspondence: ivan.contreras@udg.edu (I.C.); josep.vehi@udg.edu (J.V.)

Abstract: This study investigates how missing data samples in continuous blood glucose data affect the prediction of postprandial hypoglycemia, which is crucial for diabetes management. We analyzed the impact of missing samples at different times before meals using two datasets: virtual patient data and real patient data. The study uses six commonly used machine learning models under varying conditions of missing samples, including custom and random patterns reflective of device failures and arbitrary data loss, with different levels of data removal before mealtimes. Additionally, the study explored different interpolation techniques to counter the effects of missing data samples. The research shows that missing samples generally reduce the model performance, but random forest is more robust to missing samples. The study concludes that the adverse effects of missing samples can be mitigated by leveraging complementary and informative non-point features. Consequently, our research highlights the importance of strategically handling missing data, selecting appropriate machine learning models, and considering feature types to enhance the performance of postprandial hypoglycemia predictions, thereby improving diabetes management.

Keywords: classification; data quality; hypoglycemia prediction; machine learning; postprandial hypoglycemia; type 1 diabetes; missing data

MSC: 68T09

check for updates

Citation: Rehman, N.U.; Contreras, I.; Beneyto, A.; Vehi, J. The Impact of Missing Continuous Blood Glucose Samples on Machine Learning Models for Predicting Postprandial Hypoglycemia: An Experimental Analysis. *Mathematics* **2024**, *12*, 1567. https://doi.org/10.3390/math12101567

Academic Editor: Cristina I. Muresan

Received: 28 March 2024
Revised: 25 April 2024
Accepted: 15 May 2024
Published: 17 May 2024

1. Introduction

Diabetes is a group of metabolic disorders distinguished by high blood glucose (BG) levels [1]. Long-term complications of diabetes can include heart disease, renal failure, circulatory disorders, and nerve damage [2,3]. By 2040, an estimated 642 million people worldwide will have diabetes [4]. Diabetes is mainly classified into three types: (i) type I (T1D), (ii) type II, and (iii) gestational diabetes [5]. T1D is most commonly found in young adults under the age of 30, the symptoms of which include polyuria, thirst, constant hunger, weight loss, vision changes, and fatigue [6]. Commonly, type 2 diabetes affects adults over the age of 45 and is frequently associated with obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases [7]. Gestational diabetes, the third type of diabetes, affects pregnant women.

Despite large fluctuations in BG due to meals, exercise, stress, etc., BG concentrations should be maintained within a healthy range of 70–180 mg/dL throughout the day. To maintain normal BG levels, the body must sustain stable glucose levels and insulin production. An imbalance between BG and insulin could lead to hypoglycemic or hyperglycemic events. Hypoglycemia occurs when blood sugar levels fall below 70 mg/dL, while hyperglycemia occurs when they rise above 180 mg/dL. Hypoglycemia is further categorized into two levels: level 1 hypoglycemia occurs when BG levels drop below 70 mg/dL, and level 2 hypoglycemia occurs when they drop below 54 mg/dL [8]. This is one of the most dangerous conditions of diabetes; in severe cases, hypoglycemia can result in unconsciousness, confusion, muscle spasms, and even death [9]. Long-term complications of untreated hyperglycemia include cardiovascular disease, nerve damage (neuropathy), ketoacidosis, diabetic retinopathy, damaged nerves, or poor blood flow, which can result in serious skin infections, ulcerations, and, in severe cases, amputation [10].

The severity of hypoglycemia and hyperglycemia highlights the importance of managing diabetes properly. Therefore, diabetic patients, especially patients with T1D, use glucose monitoring devices for effective diabetes management [11]. Continuous blood glucose monitors (CGMs) read BG levels every few minutes, making them one of the best ways to keep track of BG levels. Numerous devices are available on the market, such as the Medtronic CGM, Abbott FreeStyle Libre, Dexcom CGM systems, and many others [12,13]. CGM sensors typically read the BG level every 5 min. However, these sensors may miss BG readings for several reasons. For example, missing data might occur due to a lost connection between the sensor and the recipient, such as a mobile phone or another device, due to a low battery or during a battery replacement, and, in rare cases, an interrupt signal generated by the system updating or hardware-related issues, etc. Because of these causes, BG samples are sometimes unavailable, which can significantly affect diabetes management [14].

Missing values are a common problem in all data-related fields, causing a variety of problems such as data analysis issues, performance degradation, and biased outcomes [15]. Furthermore, the pattern of missing data, the amount of missing data, and the mechanism underlying the missingness of the data all contribute to the severity of the missing values. The lack of data lowers statistical power, and lost data may result in bias estimation. Moreover, it has the potential to reduce sample representation in the dataset, which leads to imbalanced classes. Furthermore, this could complicate a study's analysis; one might not be sure whether the missing data have influenced the outcome or not. Each one of these distortions can jeopardize the significance of the trials and lead to incorrect conclusions. Researchers frequently concentrate on handling missing data, issues related to missing data, and ways to prevent or try to reduce it in clinical research [16–18]. Most researchers, however, have relied on the presumption of a complete dataset when coming to their conclusions until recently. Little attention has been paid to the general subject of missing data in the context of diabetes care.

Machine learning (ML)-based systems have significantly advanced early diagnosis and support mechanisms for diseases such as cancer, coronary artery disease, and stroke, enabling predictive, detection, and preventive measures [19]. This advancement has also benefited diabetes management [20]. However, analyzing diabetes-related data presents challenges due to their non-linear, non-normal, and complex nature [21–26]. The literature reveals a variety of ML models deployed for diabetes prediction, with artificial neural networks (ANNs) being most common [27]. ANNs are favored for their ability to learn from data autonomously, extracting features based on hidden parameters. Various ANN variants have been employed for prediction or classification tasks [28–41]. Additionally, decision trees (DTs) and random forest (RF) are widely used for their effectiveness in general classification and regression, offering simplicity, ease of implementation, and accuracy, making them attractive for diabetes management applications [29,38,42–48]. Naive Bayes (NB) models are also utilized to effectively handle unbalanced classes, operating on the principle that all features are independent [45,49,50]. Support vector machines (SVMs) with a kernel-based approach are preferred for predicting or detecting hypoglycemia for minimal data scenarios. SVMs, binary classifiers that employ the kernel function to optimally separate outputs, rely on complex data transformations to categorize input data in a multi-dimensional space, with the kernel function influencing the hyperplane's shape and decision boundaries [28,34,38,44,49,51–53]. Logistic regression (LR) is another technique for predicting outcomes with discrete labels. LR models apply a logistic function to correlate the input with the probability of one or more classes, adjusting their algorithms and solvers to suit the dataset's type and size [44,46,47,49,51]. A total of six ML models, i.e., naive Bayes (NB), random forest (RF), logistic regression (LR), artificial neural network (ANN), decision tree (DT), and support vector machine (SVM), were shortlisted for the study, based on the families of the ML models discussed by Mujahid et al. [27].

One of the most challenging issues for accurate hypoglycemia risk stratification in T1D patients is developing a robust ML model. In ML models, missing samples are commonly addressed using imputation methods [54]. These methods involve replacing the missing data points or samples with values obtained through different statistical techniques. In practice, it may be challenging to ascertain whether the assumptions underlying imputation methods (e.g., the majority of methods assume data are missing at random) are satisfied, as this can significantly impact the performance of such methods [54–56]. Given these considerations, a comprehensive experimental analysis was undertaken to investigate multiple questions related to hypoglycemia prediction in light of missing data. The major question this analysis asks is how missing BG samples impact the performance of the ML models in predicting postprandial hypoglycemia. This question can be further categorized into four sub-questions. The first question asks how well ML models perform when predicting postprandial hypoglycemia when BG samples are missing in the preprandial window. Specifically, can these models handle missing samples sufficiently effectively to make accurate predictions about hypoglycemia? The second question explores how the performance of ML models deteriorates with the increase or decrease in the amount of missing BG samples in the preprandial period. The third question focuses on determining which ML model is most effective for predicting postprandial hypoglycemia when samples are missing. The fourth question examines whether the feature impacts the performance of ML models when data are missing in the preprandial window. Finally, seven classic interpolation methods were selected, to analyze their impacts on the prediction models [57]. These questions aim to shed light on the challenges of using ML models in diabetes management and to help researchers in their decision when choosing the best ML model for hypoglycemia prediction.

In light of the aforementioned issues, the following contributions are made:

- The effect of the quantity of missing BG samples on the ability of the ML model to predict postprandial hypoglycemia.
- The impact of the position of occurrence of missing samples on the ML model for postprandial hypoglycemia prediction.
- The role of missing samples in the relevance of features for predicting hypoglycemia events.

The paper is structured in the following manner: Section 2 presents the methodology used, which includes various scenarios involving missing samples, different interpolation techniques, machine learning models, and evaluation metrics. The results of the study are briefly presented in Section 3, wherein we evaluate the performance of each model. The discussion in Section 4 not only explores the results of these findings but also discusses the challenges and limitations that were faced during our analysis, providing a clear overview of the study's scope and constraints. The concluding remarks in Section 5 concisely summarize the contributions of our study and suggest possible directions for future research.

2. Methodology

2.1. Experimental Datasets and Preprocessing

A simulator based on the Hovorka model [58] was presented with challenging and realistic scenarios for T1D patients, including a library of mixed meals, a model of circadian variability of insulin sensitivity, and also virtual patient (VP) generation. We used this simulator to generate 249 VPs for in silico testing. First, a 30-day scenario was defined with three meals: breakfast, lunch, and dinner, served at 08:00 ± 50 min, 13:00 ± 50 min, and 20:00 ± 50 min, respectively. After simulation, treatment decisions were made based on simulated self-monitoring BG measurements by a CGM sensor; with a sample time of 5 min, we obtained 288 samples per day, for a total of 8640 samples per patient. The data were preprocessed to collect all the necessary information, and patients with fewer than 15 cases of hypoglycemia events in 30 days were excluded, resulting in 210 VPs for the final analysis.

To validate the analysis, a cohort of 10 real patients with T1D was also included. The study was limited to adults, with no patients under 18. The average (±SD) monitoring period for each patient was 124 ± 26 days. To monitor glucose concentration, each patient used the CGM FreeStyle Libre system (Abbott Diabetes Care, Alameda, CA, USA) [28]. For each pump model, we customized a routine to extract the critical data into a comma-separated value (csv) file, including the dates, timestamps, insulin delivered, meals consumed, and CGM signal for each patient. After extracting the data into a CSV file, the next step was to clean and organize the data in a frame, with the relevant features as columns, to ensure data integrity.

2.2. Missing Data Scenarios

An exhaustive literature review was performed to represent real-world missing scenarios in CGM data; however, no information about the pattern of CGM missing data due to different causes was reported, and thus, we considered two distinct strategies. The selection of these specific strategies was inspired by the comprehensive findings from a study [59] which evaluated the effects of various degrees and patterns of data loss on CGM metrics' accuracy. This foundational work provided a scientific basis for our choices, aiming to simulate a broad spectrum of real-life disruptions in CGM data, from minimal to significant data loss. The chosen percentages and durations reflect common real-world challenges CGM users face, ensuring that our simulated scenarios are relevant and based on practical experiences. The first strategy produces custom scenarios, all of which are based on the presumption that the absolute worst case that could happen takes place. The second strategy is based on the assumption that real-world missing data scenarios do not follow any particular pattern; these are referred to as random scenarios. For custom scenarios, time-series data leaks were simulated just before meal ingestion. These missed values may represent network failures, hardware damage, or battery depletion of devices or sensors. On the other hand, point-in-time data losses were simulated. This may reflect intermittent device connection problems, short signal losses, or hardware errors. In this case, BG samples were deliberately removed at random time steps. In total, four different time window lengths of data loss were simulated:

1. Five minutes of missing data (removal of a single sample).
2. Half an hour of missing data (removal of six samples).
3. One hour of missing data (removal of twelve samples).
4. Two hours of missing data (removal of twenty four samples).

Based on these time window lengths of data loss, scenarios with an incremental percentage of missing data (10%, 30%, 50%, 70%, and 90%) were also included in the study. Before the feature engineering process, these scenarios were simulated by eliminating values six hours before each meal. The samples were removed so that no overlapping of missed data occurred during the procedure of removing samples, as shown in Algorithm 1. Aside from the values of BG, all other features, such as insulin on board (IOB), and rate of carbs appearance in blood glucose (RA), remained unchanged, and there were no missing data.

In the six hours before the meal, the real patient dataset displayed a baseline of 3.01% missing data on average. In addition to meal instances with missing data percentages of 10%, 30%, 50%, 70%, and 90%, meal instances with missing data that were previously present but had a higher quantity than the embedded missing data were also considered. Thus, the overall missing data per meal instance combined the embedded and already present missing data, as shown in Table 1. However, in some instances, the already-present missing data were more than the embedded missing data, so those instances remained unchanged. From the table, it can be seen that by adding 10% more data before the meal, the average percentage of missing data in an instance across the dataset increased to 12.08%; similarly, by adding 30%, 50%, 70%, and 90% more data, the average percentage of missing data per instance increased to 31.27%, 50.72%, 71.49%, and 90.21%, respectively. This was because of the instances where the already-missing data were higher than the added ones. The random process of removing data was performed 20 times for both cohorts, to achieve unbiased and accurate results. Thus, we obtained a total of 400 configurations for each of the databases described above.

Algorithm 1 Ensure no overlapping of missed data

```
Define $\tau$, the time window for potential data removal
Calculate $n$, the required number of samples to remove
Initialize $count \leftarrow 0$
for $i \leftarrow 1$ to length$(\tau)$ do
    if status$(i) \neq$ 'removed' then
        Remove sample $i$ and set status$(i) \leftarrow$ 'removed'
        $count \leftarrow count + 1$
        if $count = n$ then
            break from the loop
        end if
    end if
end for
Ensure status$(i) \neq$ 'overlap' for all $i$
Perform data integrity check
```

Table 1. Overall missing continuous blood glucose samples in real patient dataset.

Instance Type	Missing per Type	Cumulative Missing Samples	Overall Percentage Missing
Original data	17,928	17,928	3.01%
Embedded missing 10% Original > 10%	20,148 6429	26,577	12.08%
Embedded missing 30% Original > 30%	63,182 5604	68,786	31.27%
Embedded missing 50% Original > 50%	107,446 4108	111,554	50.72%
Embedded missing 70% Original > 70%	155,529 1721	157,250	71.49%
Embedded missing 90% Original > 90%	198,075 348	198,423	90.21%

Six scenarios were considered for custom scenarios or time-series data leaks; each scheme is shown in Table 2. Custom scenarios were only considered for virtual patients because the real patient dataset already contains missing data, making it impossible to analyze the anticipated scenarios' importance to calculate accurate results. Figure 1 shows the diagrammatic representation of the process used to embed missing data scenarios.

Table 2. Custom scenarios' details.

Scenario	Available BG Samples	Description
S1	72	No data removed before meal
S2	60	1 h (12 samples) removed before meal
S3	48	2 h (24 samples) removed before meal
S4	36	3 h (36 samples) removed before meal
S5	24	4 h (48 samples) removed before meal
S6	12	5 h (60 samples) removed before meal

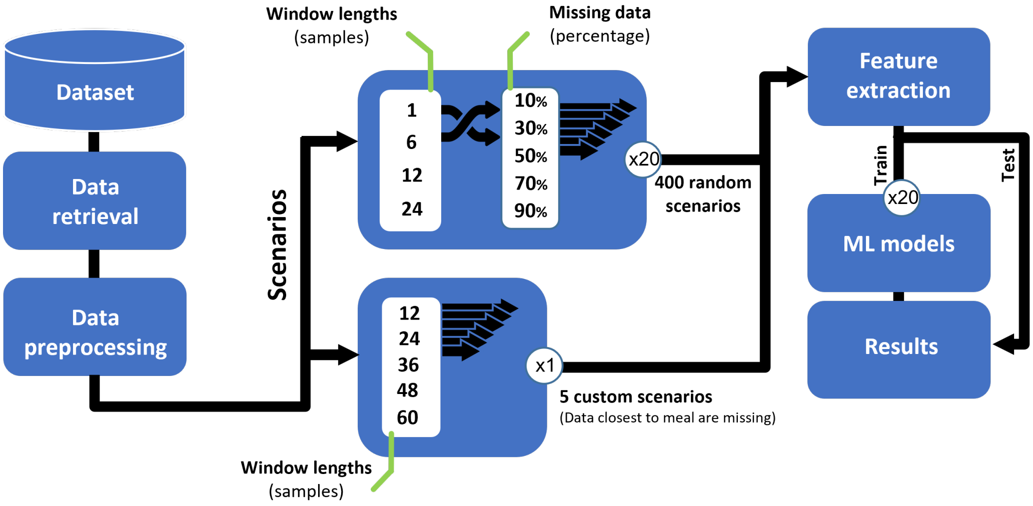

Figure 1. General methodology for embedding missing data scenarios.

2.3. Interpolation Techniques for Handling Missing Data

Several interpolation methods were explored to simulate scenarios involving missing data, each using a distinct methodology. To prevent any biased results, the root-mean-square error (RMSE) was calculated for each interpolated scenario against all the missing data samples, as shown in Table 3. The scenario with the lowest RMSE values was considered, meaning only the scenario where one sample was randomly missed was considered.

- For linear interpolation, the missing value was calculated based on the proportionate distance between the two known samples along the line. Given two known samples (x_1, y_1) and (x_2, y_2), the formula reads as follows:

$$y = y_1 + \frac{(y_2 - y_1)}{(x_2 - x_1)}(x - x_1)$$

- The nearest-neighbor interpolation method is simple to implement but may only sometimes provide accurate estimates, especially if the nearest data samples represent the missing value.

$$y = y_{\text{nearest}}$$

- The next and previous interpolation technique is intuitive and quick to apply but may introduce bias if the data samples are not evenly distributed.

$$y_{\text{missing}} = y_{\text{next or previous}}$$

- The Pchip interpolation technique is particularly useful for preserving trends in the data and avoiding overshooting.

$$y(x) = \sum_{i=1}^{n-1} \Big(a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \Big)$$

  where (x_i, y_i) are the known data samples, and n is the number of data samples. The coefficients a_i, b_i, c_i, and d_i are determined such that the slope at each data sample matches the slope of the cubic spline interpolating the data.
- The Makima interpolation method effectively reduces oscillations and improves accuracy in highly fluctuating datasets.

$$y(x) = \sum_{i=1}^{n-1} \Big(w_i^3(x - x_i)^3 + w_i^2(x - x_i)^2 + w_i(x - x_i) + 1 \Big) y_i$$

where (x_i, y_i) are the known data samples, and n is the number of data samples. The weights w_i are calculated based on the slopes of the data samples, with a weighted average used to determine the interpolation polynomial at each data sample.

- Splines are beneficial for their smoothness and accuracy, making them suitable for complex datasets with irregularities.

$$S(x) = ax^3 + bx^2 + cx + d$$

where the coefficients a, b, c, and d vary from segment to segment.

Table 3. Root-mean-square error metrics under different interpolation methods and random missing scenarios for virtual patients.

Window Length	Missing Data	Interpolation						
		Linear	Makima	Nearest	Next	Pchip	Previous	Spline
1	10%	0.39	0.40	0.96	2.10	0.41	2.15	0.45
	30%	0.43	0.43	2.26	3.02	0.44	3.09	0.51
	50%	0.52	0.47	2.84	4.40	0.50	4.48	0.60
	70%	0.96	0.65	4.25	7.19	0.70	7.39	0.93
	90%	2.05	5.03	7.04	12.34	5.01	12.89	7.24
6	10%	0.56	0.53	3.01	5.15	0.57	5.23	0.70
	30%	1.14	0.78	5.24	8.99	0.84	9.25	1.06
	50%	1.82	1.16	6.73	11.65	1.24	12.02	1.56
	70%	3.16	2.05	9.34	15.95	2.12	16.72	2.46
	90%	10.29	8.28	21.10	34.28	8.60	35.72	14.08
12	10%	1.22	0.90	5.78	9.90	0.98	10.44	1.24
	30%	2.17	1.35	7.87	13.36	1.40	14.03	1.64
	50%	3.39	2.18	10.28	17.42	2.26	18.20	2.47
	70%	5.39	3.63	13.40	22.69	3.74	23.64	4.14
	90%	9.10	6.75	19.60	31.98	7.08	33.65	9.01
24	10%	1.56	1.38	7.87	12.20	1.47	12.60	1.77
	30%	4.03	2.67	12.14	20.13	2.74	21.91	2.76
	50%	8.39	5.97	18.78	30.42	6.18	32.20	6.65
	70%	8.38	6.01	18.94	30.47	6.19	32.26	6.71
	90%	14.25	10.89	30.15	49.12	11.39	51.51	16.05

2.4. Data Features

Data preparation steps were carried out to provide data to ML algorithms. Time stamps, CGM values, meal intake details, rate of carbohydrate appearance in blood glucose (RA), and basal insulin data were all included in the raw data, which were saved in a CSV format. A total of 27 time-domain features were derived from the raw data after the data integration was completed. The input comprised six hours of BG levels before the meal, and for the output, hypoglycemia was considered in the four hours after the meal, i.e., the postprandial period. The final datasets contained imbalanced classes, so the stratified k-fold function was used to ensure an equal proportion of output classes in the training and testing datasets.

The approach used in this study had a prediction horizon (PH) of 4 h after each meal. An illustration of the meals and PHs of the virtual patient cohort is provided in Figure 2. For real patients, meals that did not respect the 4 h of the postprandial period without consuming carbohydrates were not taken into account.

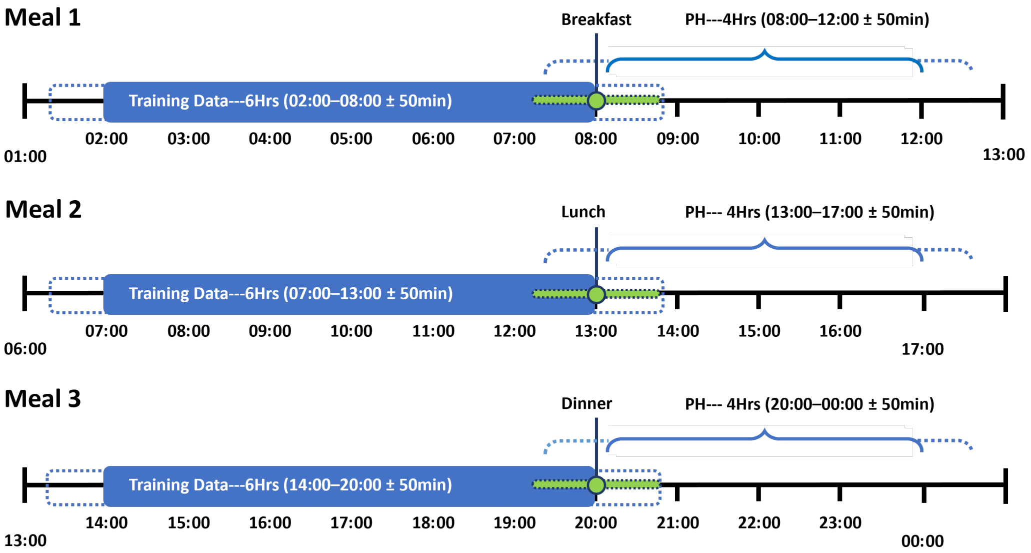

Figure 2. Pre- and postprandial time windows for virtual patients in a dataset.

In addition to features derived from BG values, features based on the insulin on board (IOB) and the rate of glucose appearance (RA) at meal intake times were also used. Class labeling was carried out four hours after the start of the meal (m_0). Three consecutive BG level readings less than or equal to 70 mg/dL were classified as class 1 (meal with hypoglycemia). Otherwise, class 0 (meal without hypoglycemia) was assigned. Below are the details of the features derived:

- **CGM at meal**: actual BG value observed at the start of meal (m_0).
- **Mean CGM meal**: mean BG levels for each hour in the last six hours before a meal, i.e., mean(BG(m_0-1), BG(m_0-2),..., BG(m_0-6))
- **Area under the curve (AUC)**: area under the curve for a BG threshold of 70 mg/dL for each hour in the last six hours before meal.
- **Rate of change (ROC)**: BG rate of change over the last 30 min with five min intervals, i.e., $(\mathrm{BG}(m_0) - \mathrm{BG}(m_0 - 1))/5$.
- **Low blood glucose index (LBGI)**: low blood glucose index in the last four hours and six hours
- **High blood glucose index (HBGI)**: high blood glucose index in the last four hours and six hours.
- **Difference Δ in CGM values**: the difference between the current CGM observation and the ones observed 30 min and 60 min earlier.
- **Insulin on board (IOB)**: Insulin on board at start of meal (m_0).
- **Rate of carbohydrate appearance in blood glucose (RA)**: rate of carbohydrate appearance in the blood glucose at the start of the meal (m_0).

2.5. Machine Learning Models

In this study, we aimed to predict postprandial hypoglycemia using ML-based systems, particularly examining the impact of missing CGM data on the predictive performance of these models. Considering the binary classification challenge and the importance of handling missing data effectively, we explored various ML models, as detailed in Section 1. The array of models analyzed in our research included the following:

- A multilayer perceptron-based artificial neural network (ANN) with a configuration of three layers, each consisting of 13 nodes, employing the Adam optimization algorithm. This neural network model was designed to capture complex patterns in the dataset.
- A support vector machine (SVM) utilizing a radial basis function (RBF) kernel to adeptly address non-linear patterns within the dataset.

- Logistic regression (LR) with a Newton-CG solver, which provides probability estimates in a binary classification context.
- Random forest (RF), an ensemble method combining multiple decision trees to bolster prediction accuracy and mitigate overfitting.
- Decision trees (DTs) are highly beneficial for healthcare applications, as they provide clear and interpretable information in the decision-making process.

The Supplementary Materials present detailed mathematical equations and hyperparameter tuning for each model, providing an in-depth technical reference for our modeling approach.

Our study utilized a k-fold cross-validation methodology for both the training and testing phases. This method, applied across 20 iterations, ensures a thorough and unbiased evaluation by systematically rotating the dataset through training and validation roles, thus maximizing the use of available data for a comprehensive assessment. The results are reported as mean values obtained from these iterations, presenting a thorough and detailed perspective of each model's performance in predicting postprandial hypoglycemia with missing CGM samples. For the analysis of the ML model under interpolation techniques, RF, ANN, and LR were considered for the study.

2.6. Technical Specifications and Performance Metrics

A Matlab script was used for all data-related operations, such as raw data preprocessing, missing data scenarios, and data features. A Python-based program with libraries such as Scikit-learn, Numpy, Pandas, and Matplotlib was developed to program prediction models and compute performance metrics [60].

As discussed in Section 2.4, the final datasets had uneven classifications, necessitating the use of the stratified k-fold function to guarantee that the same proportion of classes appeared in both the training and testing datasets. Although the accuracy produced by this train–test split was high, it was not the ideal metric to judge a model's efficacy, due to the implicit bias in the dataset. Therefore, the Matthews correlation coefficient (MCC), which determines the correlation between the true classes and the predicted labels, was used to evaluate the performance of the implemented models. The MCC accepts values between -1 and 1. A classification value of -1 shows that the model predicts the positive and negative of the target variable oppositely. If the MCC is set to 0, the model predicts positive and negative classes at random. The model is ideal when the value is 1. The MCC is calculated using Equation (1), as follows:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{1}$$

where TP denotes a true positive (a positive class is predicted as a positive class); TN denotes a true negative (a positive class is predicted as a negative class); FP denotes a false positive (a negative class is predicted as a positive class); and FN denotes a false negative (a positive class is predicted as a negative class).

Furthermore, a dataset contains numerous potential variables and features; thus, feature selection becomes crucial. By analyzing irrelevant or less significant features that contribute little to the target variable, classification performance and accuracy would be improved [61,62]. This study evaluates the significance of features in light of the missing data. After embedding each missing data scenario, a statistical analysis was conducted to select the best features [63]. In datasets with missing data, the F-score for each feature against the target variable was calculated using an analysis of variance (ANOVA) to determine the top features based on their relationship with the target variable.

3. Results

3.1. Random Missing Scenario

The original datasets were used to calculate the baseline MCC values for the virtual and real patients; in other words, the real patient data and the virtual patient data had their original datasets with no missing or additional data added. This made it possible for us to analyze and compare the performance of ML models and the impact of missing data on the models used. For example, with no alterations in the original datasets, the implemented random forest (RF) model outperformed other implemented ML models in terms of MCC, whereas the naive Bayes (NB) model and the decision tree (DT) model underperformed in the cases of datasets from virtual and real patients, respectively.

The original datasets were introduced to missing data scenarios, i.e., different window lengths and missing data percentages, and the results of each iteration for simulated data and real data are summarized in Tables 4 and 5, respectively. As the BG values were missed randomly from the preprandial samples, the features calculated from them were heavily influenced, and the RA and IOB features contributed to the learning process, which was insufficient, as seen in the tables. The tables show that the MCC values decreased as the amount of missing data increased, indicating that the performance of the ML model deteriorated.

Table 4. Matthews correlation coefficient (MCC) values under random missing scenarios for virtual patients (NB: naive Bayes, RF: random forest, LR: logistic regression, ANN: artificial neural network, DT: decision tree, SVM: support vector machine).

Window Length	Missing Data	Algorithms					
		NB	**RF**	**LR**	**ANN**	**DT**	**SVM**
Baseline Results		0.59	0.74	0.63	0.66	0.66	0.63
1	10%	0.59	0.73	0.62	0.63	0.64	0.62
	30%	0.57	0.71	0.60	0.61	0.60	0.61
	50%	0.54	0.68	0.56	0.58	0.57	0.57
	70%	0.51	0.67	0.54	0.56	0.56	0.54
	90%	0.52	0.66	0.52	0.53	0.55	0.52
6	10%	0.51	0.68	0.55	0.58	0.58	0.56
	30%	0.51	0.67	0.54	0.56	0.57	0.55
	50%	0.51	0.67	0.52	0.55	0.57	0.53
	70%	0.50	0.67	0.51	0.53	0.56	0.52
	90%	0.50	0.65	0.50	0.52	0.54	0.50
12	10%	0.51	0.68	0.54	0.57	0.58	0.55
	30%	0.51	0.67	0.52	0.55	0.57	0.54
	50%	0.50	0.67	0.51	0.54	0.57	0.52
	70%	0.50	0.66	0.50	0.53	0.56	0.51
	90%	0.49	0.65	0.49	0.52	0.55	0.50
24	10%	0.52	0.68	0.57	0.59	0.59	0.57
	30%	0.51	0.67	0.54	0.55	0.57	0.54
	50%	0.51	0.66	0.52	0.54	0.56	0.52
	70%	0.50	0.66	0.51	0.54	0.56	0.52
	90%	0.50	0.65	0.50	0.54	0.56	0.52

From the tables, it can be deduced that the implemented RF model performed better than the other models, in terms of MCC. For both datasets, i.e., virtual and real, the MCC values of the RF model showed better results, despite the increment in missing data, compared to other implemented models. While analyzing the data, it is essential to remember that the ML model was most affected by missing data in the real patient dataset when it occurred randomly across the preprandial with the window length of a single sample. Due to the low correlation within the dataset, the ML model could not effectively detect the missing samples when the missing data were spread out across the time period. This contrasts with the virtual patient dataset, which has a high correlation between the variables.

Table 5. Matthews correlation coefficient (MCC) values under random missing scenarios for real patients (NB: naive Bayes, RF: random forest, LR: logistic regression, ANN: artificial neural network, DT: decision tree, SVM: support vector machine).

Window Length	Missing Data	Algorithms					
		NB	**RF**	**LR**	**ANN**	**DT**	**SVM**
Baseline Results		0.40	0.49	0.46	0.44	0.31	0.49
1	10%	0.32	0.47	0.41	0.36	0.30	0.41
	30%	0.30	0.43	0.39	0.33	0.26	0.38
	50%	0.29	0.42	0.37	0.33	0.25	0.38
	70%	0.28	0.39	0.35	0.31	0.22	0.35
	90%	0.23	0.33	0.25	0.22	0.19	0.27
6	10%	0.32	0.48	0.41	0.40	0.30	0.43
	30%	0.30	0.47	0.40	0.37	0.30	0.41
	50%	0.30	0.46	0.39	0.36	0.29	0.40
	70%	0.29	0.45	0.39	0.34	0.27	0.39
	90%	0.29	0.45	0.38	0.34	0.27	0.39
12	10%	0.34	0.48	0.43	0.42	0.30	0.44
	30%	0.32	0.48	0.42	0.40	0.30	0.43
	50%	0.31	0.47	0.41	0.38	0.30	0.42
	70%	0.30	0.46	0.40	0.37	0.29	0.41
	90%	0.29	0.46	0.39	0.36	0.28	0.40
24	10%	0.34	0.48	0.42	0.41	0.30	0.44
	30%	0.34	0.48	0.42	0.41	0.30	0.44
	50%	0.32	0.48	0.41	0.40	0.30	0.42
	70%	0.31	0.47	0.41	0.39	0.29	0.42
	90%	0.23	0.33	0.25	0.21	0.18	0.27

To make it simple to assess the performance trend of each ML model and to test its robustness against missing data, a percentage decrease for each scenario was calculated. Figure 3 displays the percentage decrease for the random missing scenarios with various window lengths embedded in the virtual patient dataset. The most and least affected ML models by the missing data in the case of virtual patient data are listed in Table 6. With the percentage decrease analysis, it was concluded that, when predicting postprandial hypoglycemia with random missing data, overall, the implemented RF performed better than other ML models regarding missing data. On the other hand, in only two scenarios, when increasing the missing data from 0% to 10% and 30% with a window length of 1 sample, NB had the smallest percentage reduction compared to the other five ML models.

RF performed better than other models, followed by NB (only in two scenarios). Contrary to the least affected model, it was also noted that LR was generally inappropriate for usage, since a rise in missing data resulted in model degeneration. During the analysis, it was interesting to note that a small percentage of missing data points, i.e., 10%, that occurred in consecutive samples over the six hours before the meal had a more significant impact on the ML models than a random missing single sample. The ML models could detect the pattern easily when a single sample at random was removed over the preprandial window, due to the strong correlations within variables in the simulated data.

Figure 4 displays a percentage decrease graph for the real patients. The implemented RF, on average, was least affected by the percentage of missing data, followed by DT in some scenarios, as shown in Table 7. At the same time, the implemented NB was the most affected by missing data. It is important to remember that even though these percentage decreases for algorithms were modest compared to other models, their MCC values were notably low, making it unwise to use them in missing data situations. One must consider the percentage decrease and MCC values when determining a robust classifier model. Hence, it is evident from the analysis that the implemented RF model performed well in the case of missing data.

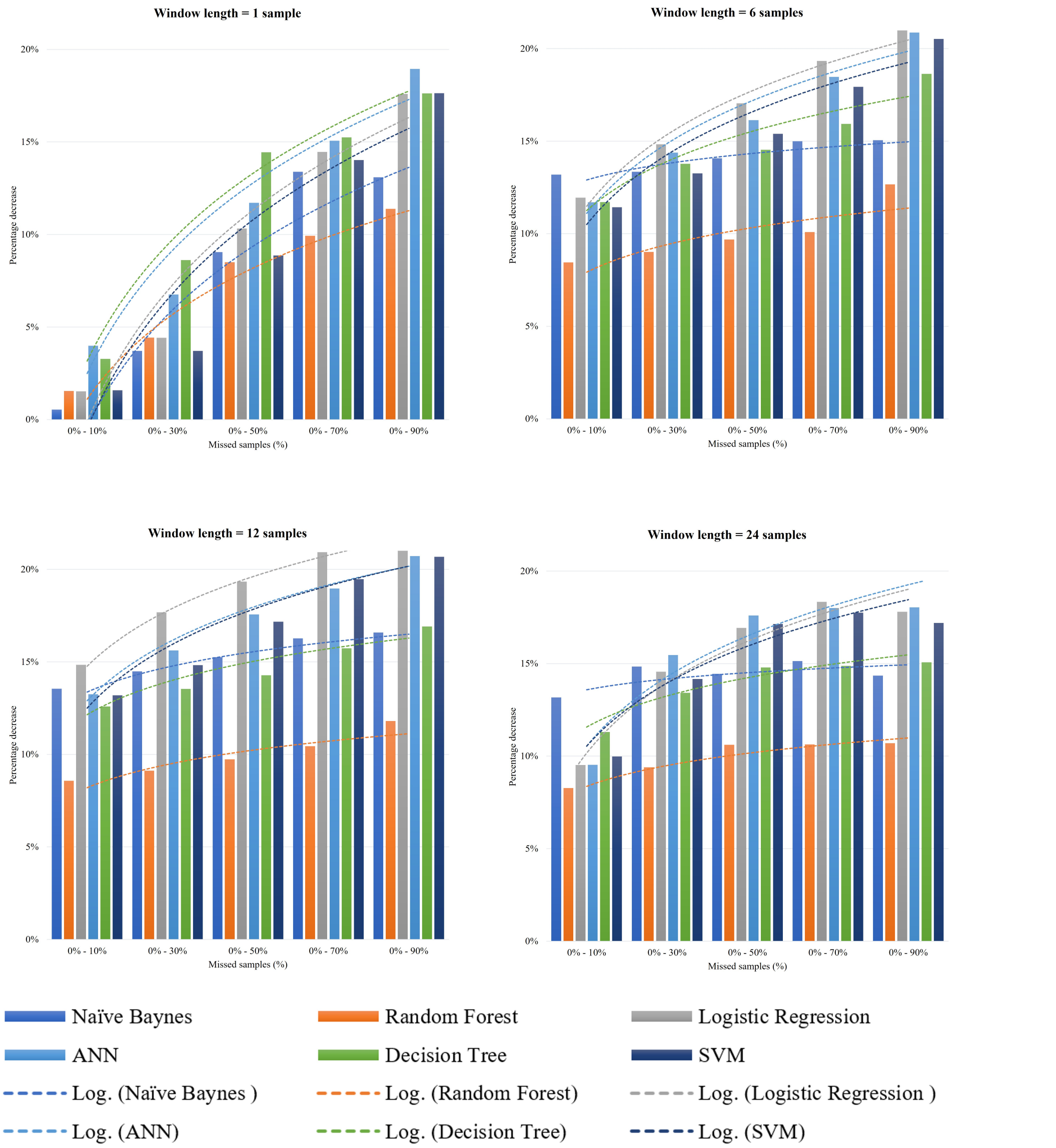

Figure 3. Percentage decrease in Matthews correlation coefficient (MCC) values with different window lengths for virtual patients.

Table 6. Virtual patients: summary of ML models affected by missing data in a random missing data scenario (NB: naive Bayes, RF: random forest, LR: logistic regression, ANN: artificial neural network, DT: decision tree).

Model Detail		**Missing Data Range**				
ML Models	**Window Length**	**0–10%**	**0–30%**	**0–50%**	**0–70%**	**0–90%**
Least affected	1 sample	NB	NB	RF	RF	RF
	6 consecutive samples	RF	RF	RF	RF	RF
	12 consecutive samples	RF	RF	RF	RF	RF
	24 consecutive samples	RF	RF	RF	RF	RF
Most affected	1 sample	ANN	DT	DT	DT	ANN
	6 consecutive samples	NB	LR	LR	LR	LR
	12 consecutive samples	LR	LR	LR	LR	LR
	24 consecutive samples	NB	ANN	ANN	LR	ANN

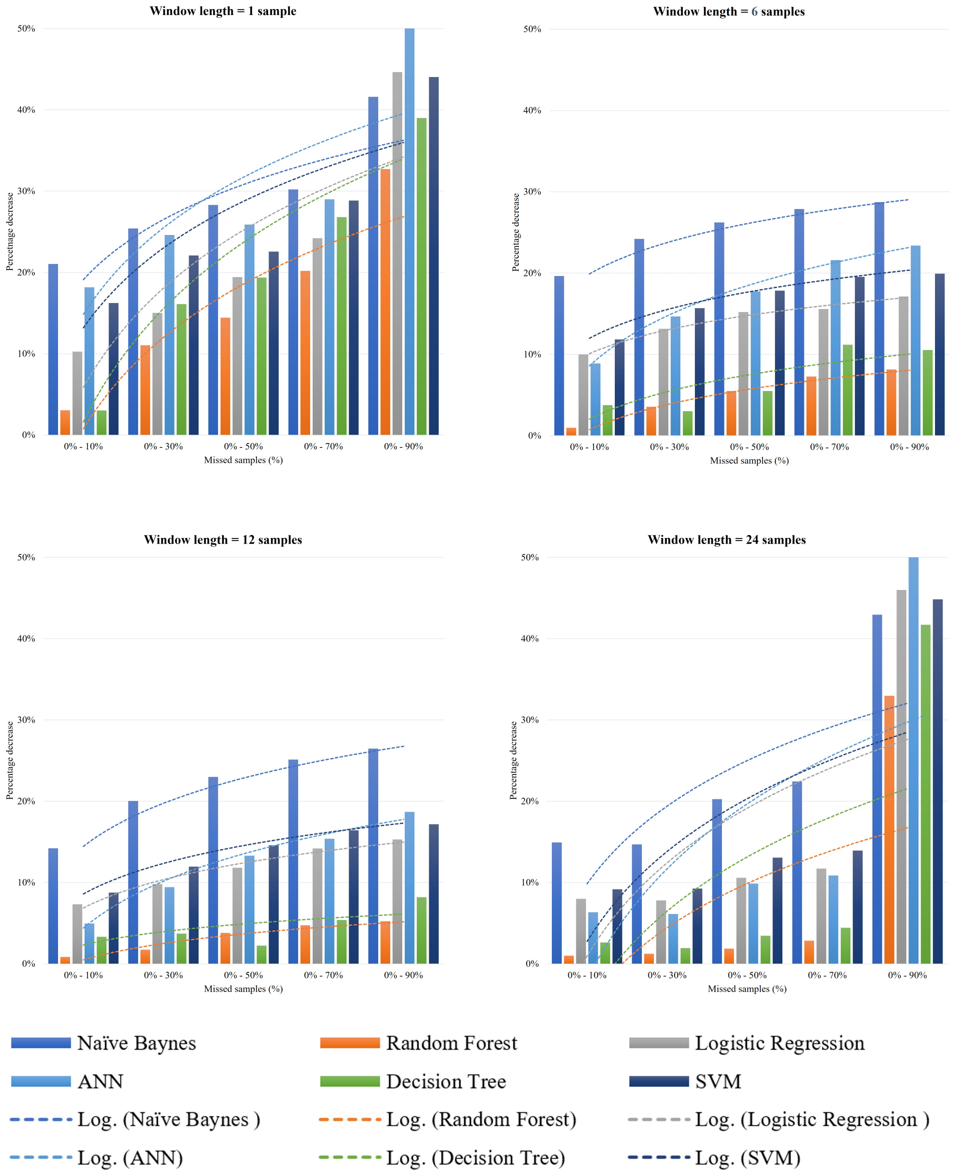

Figure 4. Percentage decrease in Matthews correlation coefficient (MCC) values with different window lengths for real patients.

Table 7. Real patients: summary of ML models affected by missing data in a random missing data scenario (NB: naive Bayes, RF: random forest, ANN: artificial neural network, DT: decision tree).

Model Detail		**Missing Data Range**				
ML Models	**Window Length**	**0–10%**	**0–30%**	**0–50%**	**0–70%**	**0–90%**
Least affected	1 sample	RF	RF	RF	RF	RF
	6 consecutive samples	RF	DT	DT	RF	RF
	12 consecutive samples	RF	RF	DT	RF	RF
	24 consecutive samples	RF	RF	RF	RF	RF
Most affected	1 sample	NB	NB	NB	NB	ANN
	6 consecutive samples	NB	NB	NB	NB	NB
	12 consecutive samples	NB	NB	NB	NB	NB
	24 consecutive samples	NB	NB	NB	NB	ANN

3.2. Custom Missing Scenario

Figure 5 compares the MCC of various implemented ML models to custom missing scenario case studies. The y-axis displays the mean MCC values obtained from the k-fold iterations, and the x-axis displays the number of BG samples available before the meal. The graphs demonstrate that the ML model degraded as the amount of missing data rose and the quantity of BG samples fell. It should be noted that in addition to features derived from BG values there were also features derived from RA and IOB values. However, as illustrated in the figure, these had no significant impact on ML model learning.

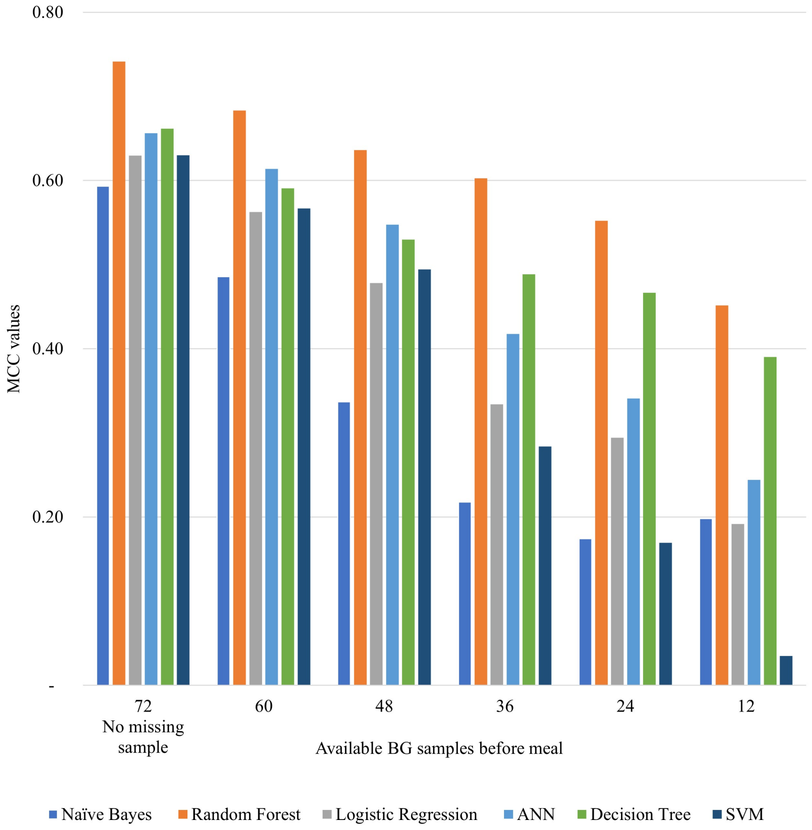

Figure 5. Mean Matthews correlation coefficient (MCC) values against custom missing scenarios.

The graph demonstrates that the results from the 60 and 48 available data samples were still acceptable, compared to other scenarios. The high MCC values resulted from the feature chosen for analysis, which involved calculating the hourly mean and smoothing out the missing data. This led to the conclusion that choosing the appropriate features can mitigate the position of missing data. However, the results of 12 data samples show that the ML models' performance was still negatively impacted by the rise in the number of missing data points.

During the percentage decrease analysis for the custom scenarios, each hourly missed data point was compared to a baseline with no missing data. The same trend was observed; ML model performance dropped significantly as the missing data increased. Overall, despite the increase in missing data in each case, the performance of the RF model implemented in this study tended to be better than that of other ML models. For example, when only one hour was missed before the meal, the implemented ANN showed better results, with a 6.46% decrease in MCC values from baseline, followed by RF with a 7.85% drop. The implemented models, i.e., NB and SVM, were the most affected ML models, with average decreases of 52.39% and 50.81%, respectively, as shown in Figure 6.

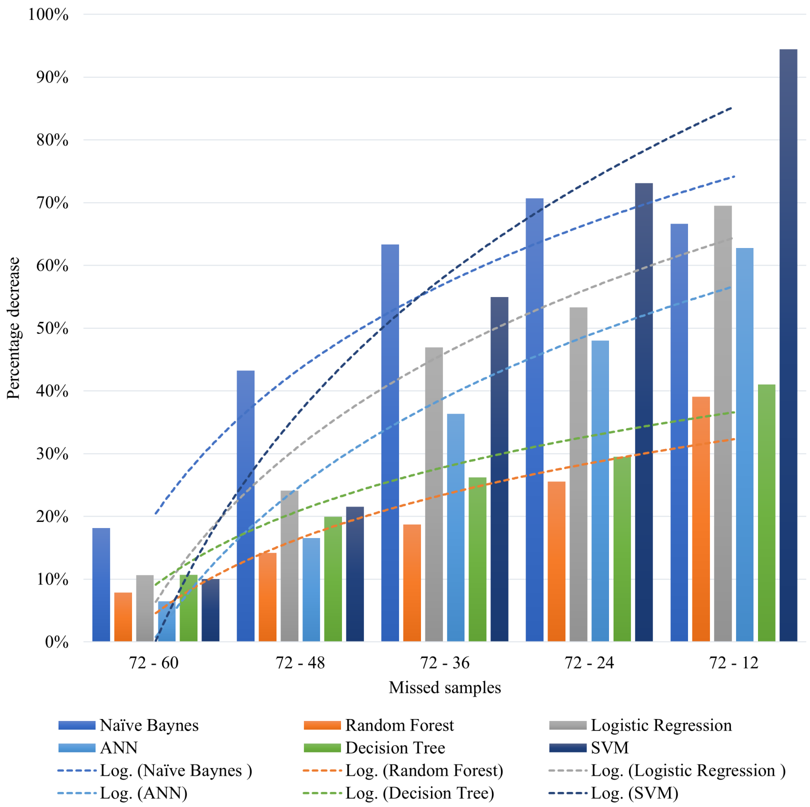

Figure 6. Percentage decrease in Matthews correlation coefficient (MCC) values in custom missing scenario.

3.3. Impact of Interpolation Techniques

In this study, three ML algorithms, RF, LR, and ANN, were subjected to a series of interpolation techniques to address missing data scenarios. Table 8 shows the resultant performance metrics of these models under varying data samples' missing percentages.

Table 8. Matthews correlation coefficient (MCC) values under random missing scenarios for virtual patients after interpolation (RF: random forest, ANN: artificial neural network).

Missing Data			**10%**			**30%**			**50%**			**70%**			**90%**	
ML Models		**RF**	**ANN**	**LR**	**RF**	**ANN**	**LR**	**RF**	**ANN**	**LR**	RF	**ANN**	**LR**	**RF**	**ANN**	**LR**
Baseline results		0.73	0.63	0.62	0.71	0.61	0.60	0.68	0.58	0.56	0.67	0.56	0.54	0.66	0.53	0.52
Interpolation	Linear	0.68	0.58	0.57	0.68	0.58	0.57	0.67	0.58	0.57	0.67	0.59 *	0.56	0.66	0.56	0.55
	Makima	0.68	0.58	0.57	0.68	0.58	0.57	0.68	0.57	0.57	0.68	0.58	0.56	0.67	0.55	0.55
	Nearest	0.67	0.58	0.57	0.67	0.58	0.57	0.67	0.57	0.57	0.67	0.57	0.56	0.67	0.56	0.55
	Next	0.67	0.58	0.57	0.67	0.58	0.57	0.68	0.58	0.57	0.67	0.59 *	0.56	0.66	0.57	0.56
	Pchip	0.67	0.58	0.57	0.67	0.58	0.57	0.68	0.58	0.57	0.68	0.58	0.56	0.67	0.55	0.55
	Previous	0.67	0.58	0.57	0.67	0.58	0.56	0.67	0.57	0.56	0.67	0.57	0.55	0.63	0.52	0.50
	Spline	0.67	0.58	0.57 *	0.67	0.58	0.57	0.68	0.59 *	0.56	0.68	0.58	0.56	0.69 *	0.56	0.53

* Indicates the highest MCC value achieved in each category.

As shown in Table 4, the baseline results for MCC values for RF, ANN, and LR in the absence of missing data were established as 0.74, 0.66, and 0.63, respectively. Even though different interpolation approaches were used, these standards have still not been surpassed or met. Notably, in an extreme case where 90% of the data samples were missing, the spline interpolation technique marginally elevated the RF model's performance to 0.69, a figure modestly higher than the 0.66 observed without interpolation in Table 4. The ANN model, however, attained a maximum MCC of 0.59, even with 50% of the data missing, facilitated by the linear and next interpolations—yet, this remains below its initial baseline results. For LR, the highest achieved MCC with spline interpolation was 0.571.

For the RF model, both linear and spline interpolations demonstrated respectable usefulness, particularly as the proportion of missing data increased, with spline having a slight edge in the 90% missing data scenario. It is crucial to acknowledge that the decrease in RF's performance as missing data increased was comparatively less prominent than that of ANN and LR across all evaluated interpolation techniques.

The linear interpolation method conferred a degree of robustness to the ANN model, maintaining an MCC level above 0.56 despite up to 70% missing data. Meanwhile, other methods exhibited only marginal performance variations compared to linear and did not provide any substantial enhancement.

For the LR model, performance remained consistent across different interpolation methods and levels of missing data, typically achieving MCC values around 0.56 to 0.57. This uniformity suggests that LR's effectiveness was relatively stable, regardless of the interpolation method used, indicating resilience to varying degrees of missing data. Unlike the RF and ANN models, complex interpolation methods do not significantly enhance LR's performance.

3.4. Impact of Missing Data on Feature Importance

This study examined feature selection using the F-measure of ANOVA for a classification problem involving numerical inputs and categorical outputs, determining the relationship between each feature and the outcome variable. Figure 7 illustrates the impact of missing data on the features extracted from the dataset after each missing data scenario was embedded and the F test score for each extracted feature against the output variable was calculated.

According to the heat map analysis, increased missing data significantly impacted features dependent on BG values. In addition, it is essential to note that in random missing scenarios, the BG values recorded at the meal and the mean BG values recorded before the meal were more important than other features. However, their significance decreased, compared to other features, as the number of missing data increased for both datasets, i.e., real and virtual patients. The heat map supports the findings of prior experiments illustrating relationships between features and highlighting trends in missing data, which, in this case, indicates a decline in the efficiency of postprandial hypoglycemia prediction.

Features such as the AUC in both datasets showed different trends, i.e., the real dataset showed a higher correlation with the target variable; however, in the virtual dataset, there was little to no correlation with the target variable. Furthermore, if we look at a particular dataset, we observe that the trends in features in both datasets are coherent, decreasing from left to right. Similarly, features based on the ROC showed similar trends: low correlation in the virtual dataset and high correlation in the real dataset. After conducting an in-depth feature analysis, it was determined that the difference in preprandial hypoglycemia between the two datasets was the primary cause of this trend in both of the features; more specifically, the real patients dataset had a greater number of instances of preprandial hypoglycemia than the virtual patients dataset. Consequently, the effect of missing data on the feature in the virtual dataset had a more significant bearing than in the dataset containing information about real patients. Nevertheless, the importance of features in both datasets degraded as the missing data increased. The same scenarios were observed in LBGI, where the feature importance was lowered, compared to other data features, with the increment in missing data within each dataset. The reason behind this was probably the higher coefficient of variance (CV) of the real patient's dataset (46.41), compared to the virtual patient's dataset (39.01).

Moreover, the IOB in the real patient's dataset did not correlate with the target variable, but with the increase in the missing data, meaning the significance was slightly elevated. However, in virtual patients, the IOB showed a significant elevation in importance as the missing data increased. Therefore, it is concluded that, based on the correlation between variables within the dataset, the significance of features may vary with the amount of missing data. It was found that missing data played a similar role in significantly reducing the importance of features in custom scenarios. However, the significance of IOB and RA rose as the missing data increased. Despite this, the performance of the ML model continued to deteriorate, as evident from earlier in Figure 5.

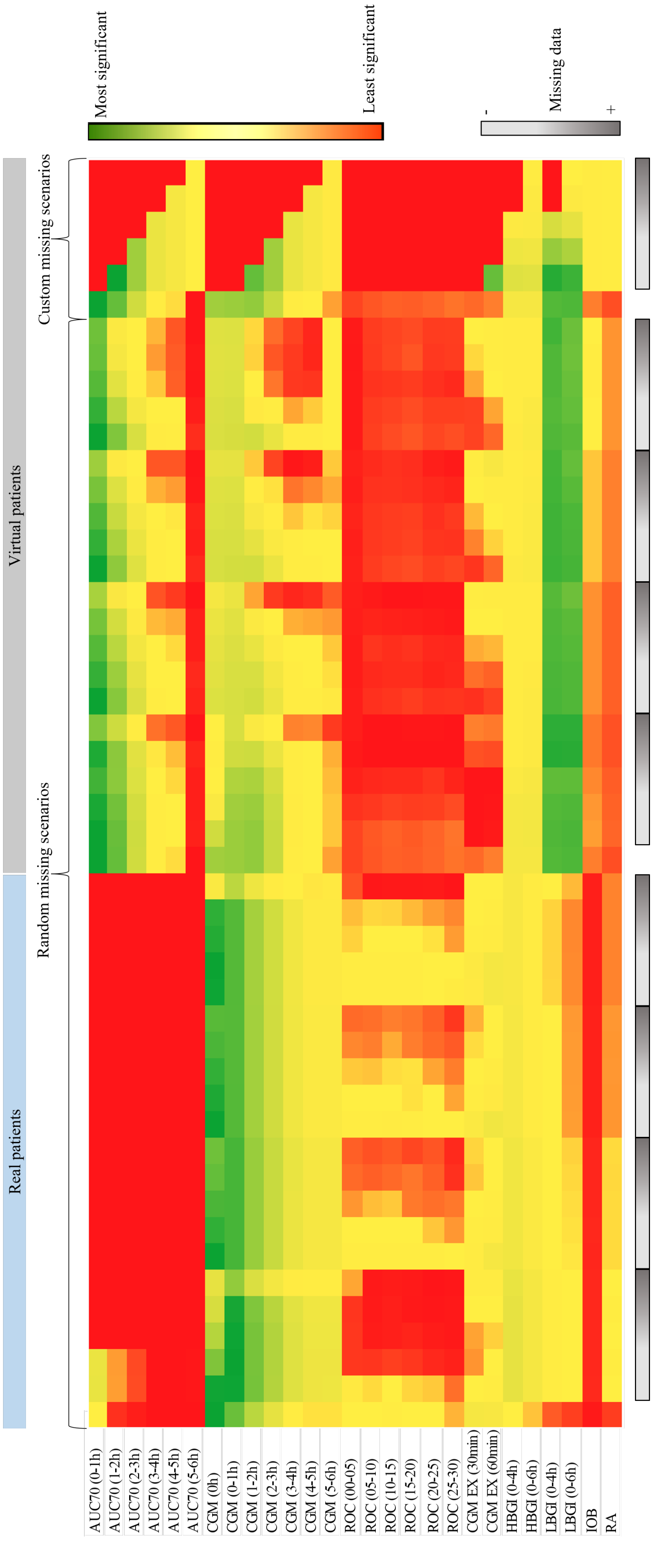

Figure 7. Heat map representing the impact of missing data on virtual and real patients under different missing scenarios.

4. Discussion

The present study conducted an exhaustive investigation into the relationship between the absence of CGM samples and the performance of ML models in predicting postprandial hypoglycemia. The results of our study demonstrate that the absence of CGM samples consistently impaired the effectiveness of ML models, confirming the crucial significance of having complete data for the precise prediction of postprandial hypoglycemia. Among the six machine learning models assessed (NB, RF, LR, ANN, DT, and SVM), the RF model exhibited better performance in the face of missing data scenarios, surpassing other models in performance across different patterns and quantities of missing data.

In conventional research, researchers often use basic imputation techniques to address missing data, assuming that the missing samples are random or without conducting a comprehensive analysis of the pattern or magnitude of the missing data. Although this approach is practical, it fails to consider the subtle influence that various missing data scenarios can have on the performance of ML models, as demonstrated by the results of our study.

The decline in model performance as the amount of missing CGM samples increased highlights the complex difficulties in accurately predicting postprandial hypoglycemia using CGM devices. The reliability of hypoglycemia predictions can be directly affected by device-related issues, such as sensor failures or connectivity problems, which can result in data loss. The findings of this study regarding the performance of models in the presence of missing samples are crucial for clinicians and patients who depend on ML-based tools for managing diabetes. These findings emphasize the importance of conducting thorough data quality assessments and creating more robust machine learning models.

By eliminating values six hours before each meal, we introduced a method pivotal for examining the impact of preprandial glucose levels on managing postprandial outcomes, notably hypoglycemia. Our study delved into how the chosen six-hour window affected the prediction outcomes of machine learning models, by simulating various extents of data removal before meals, ranging from one to five hours. These simulations involved different numbers of available BG samples post-data removal. Our findings indicated that while the machine learning models could accommodate minor data loss (e.g., one hour before meals), their performance significantly degraded with more extensive data loss (where four and five hours of data were removed). This gradient effect substantiated the hypothesis that the proximity of data loss to mealtime critically influences model performance.

Our study applied various interpolation techniques to address missing data in the six-hour preprandial window to fill missing samples in the BG. The methods used included linear interpolation, which is known for its efficiency in linearly related datasets but is potentially misleading when significant data are missing. Nearest-neighbor interpolation, which preserved local trends, led to inaccuracies in clustered or varied samples. Additionally, next and previous interpolations offered reasonable estimates but were prone to bias in non-uniformly distributed or non-linear samples. The Pchip method, while preserving shape and monotonicity, sometimes faltered when cubic polynomials did not align well with the data relationships. Makima interpolation, designed to minimize oscillations, and spline interpolation, valued for its smoothness, can be misleading when rapid changes or low-degree polynomials do not adequately represent underlying trends. Upon interpolating the missing BG data, we computed features based on hourly mean values of the BG signal. Despite the high RMSE observed in the interpolation process, as shown in Table 3, the mean values of these hourly features remained consistent with those computed before the interpolation. This consistency in the features indicates that although not accurate at every interpolated point, the methods managed to preserve the statistical properties crucial for feature generation. As a result, when these features were used in machine learning models, both interpolated and non-interpolated datasets yielded comparable predictive outcomes. This illustrates the resilience of mean-based features in maintaining critical data characteristics, even in the presence of interpolation-induced errors. In conclusion, while interpolation techniques are essential for handling missing data, they also present potential pitfalls. Researchers must recognize these risks and critically assess their assumptions and the characteristics of their specific datasets to prevent potential biases and misinterpretations. This careful consideration is necessary to ensure the reliability of data-driven insights in clinical research settings.

The study also demonstrated that the negative consequences of the missing samples could be alleviated by utilizing complementary and informative non-point features, indicating a potential approach to improve the robustness of the model. The varying influence of missing samples on the significance of features underscores the necessity for deliberate feature selection when creating ML models for hypoglycemia prediction. Furthermore, the findings of the study prompt thought-provoking inquiries regarding the characteristics of missing samples within the datasets. The present study examined custom and random missing data scenarios that replicated real-world situations, including device malfunctions or arbitrary data loss. These scenarios aimed to comprehensively comprehend the potential impact of such incidents on ML predictions within clinical settings. The varying impacts of custom and random missing data scenarios on the performance of models highlight the necessity for more advanced data-handling approaches that consider the characteristics of the missing information.

However, it is essential to acknowledge the limitations and challenges inherent in our research methodology. One limitation is the unavailability of real patient datasets with no missing data, which prompted the use of the VP dataset in our study. In real-life scenarios, information about the specific faults or events that produce certain types of missing CGM samples is also challenging to obtain. Therefore, our use of custom and random missing data scenarios aimed to simulate real-world conditions as closely as possible. Furthermore, our study primarily focused on CGM values due to their susceptibility to missing data, as this aspect reflected real-world scenarios, where data loss occurs due to CGM device faults or any other unforeseen circumstances. Nevertheless, we recognize that additional physiological markers or lifestyle factors may substantially influence the prediction of postprandial hypoglycemia. Although the current focus is on a specific aspect, including more variables in the analysis could provide more comprehensive knowledge of how missing data affect the performance of model predictions and the interaction between different factors with postprandial hypoglycemia prediction.

Our study focused on varying CGM values while keeping other features unchanged, to isolate the effects on model performance. However, this approach may have only partially captured the complexity of real-world scenarios, where multiple factors affect model performance simultaneously. To enhance the robustness and generalizability of our findings, future research should consider simulations incorporating variations in environmental conditions, patient activity levels, and other physiological variables, alongside CGM values. This multi-factorial approach could provide deeper insights into the models' sensitivity and adaptability, which is crucial for developing more effective decision support systems. Employing advanced machine learning techniques to handle high-dimensional data with varied parameters could be an interesting study to see the models' performances and their real-world applicability. By expanding the scope of variable parameters, future studies can aim to develop systems that are robust against the complexities of real-world diabetes management.

The selection of six ML models in our study was based on prior research and their established efficacy in predictive modeling tasks, which provided insights into their performance in similar healthcare applications. While the models utilized in this study offered a comprehensive representation of different ML approaches, other models or ensemble techniques could yield different results and may warrant exploration in future studies. Additionally, our study did not include an analysis of different imputation techniques due to the extensive nature of such analysis. This limitation prevented us from exploring potential methods of effectively handling missing data. Future research should incorporate various imputation techniques to determine their impact on model performance and provide more comprehensive insights into handling missing data in predictive modeling tasks.

Future research should prioritize the development of sophisticated imputation techniques to effectively mitigate the impact of missing samples on the performance of ML models. Investigating methodologies such as multiple imputations or deep-learning-based approaches may present more advanced strategies for addressing missing samples, thereby enhancing the performance and dependability of models. Furthermore, researching the influence of missing data on various medical predictions can enhance our comprehension of its ramifications in healthcare contexts. Additional investigation could also cultivate ML models specifically tailored to manage low-data-density datasets. These models have the potential to provide more dependable forecasts in situations where data loss is unavoidable, thereby improving the effectiveness of ML in clinical decision-making. Our study underscores the necessity for further investigation into feature engineering and model selection strategies capable of enduring the difficulties posed by missing data. Identifying features that maintain their predictive efficacy, even in datasets with incomplete information, can enhance the robustness of ML models, enabling them to generate accurate predictions even under suboptimal data conditions.

5. Conclusions

The impact of missing data on ML models and features is discussed in this study, with impact depending on the amount, length, and occurrence of missing data. The study uses virtual and real patients to examine the trends of missing data on six different ML models. It is worth noting that the RF produces better MCC results despite missing data. It should be noted that the position of the missing data in the dataset and the amount of missing data contribute to the degradation of the ML model. Furthermore, this study uses seven different interpolation techniques to address and evaluate the handling of missing data. Although these methods may introduce biases at some point, the overall integrity of the dataset is not compromised in any way. It was also observed that although individual samples post-interpolation exhibited high RMSE values, the incorporation of features, including mean values, negated this effect, allowing models to predict the output closely. However, the performance of these models, while near, remained within baseline results. Data scientists are advised to consider that even when there are missing data points in the dataset, better results can still be obtained by selecting the right ML model and features that smooth out the effect of missing data. Different interpolation and imputation techniques can be explored for future work to overcome the missing data for better results.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math12101567/s1.

Author Contributions: Conceptualization, N.U.R., I.C., A.B. and J.V.; data curation, N.U.R. and A.B.; formal analysis, N.U.R., I.C. and A.B.; funding acquisition, I.C. and J.V.; investigation, N.U.R. and I.C.; methodology, N.U.R. and I.C.; project administration, J.V.; resources, I.C., A.B. and J.V.; supervision, I.C., A.B. and J.V.; validation, N.U.R. and I.C.; writing—original draft, N.U.R. and I.C.; writing—review and editing, I.C. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Spanish Ministry of Science and Innovation: PID2020-117171RA-I00 and PID2022-137723OB-C22; MCIN: MCIN/AEI/10.13039/501100011033; Government of Catalonia: 2017SGR1551—2021 FI_B 00896.

Data Availability Statement: In this article's references, all datasets examined are listed, and the data can be provided upon request.

Conflicts of Interest: There are no potential conflicts of interest that need to be disclosed by any of the authors listed in the manuscript.

References

1. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **2013**, *37*, S81–S90. [CrossRef] [PubMed]
2. Krasteva, A.; Panov, V.; Krasteva, A.; Kisselova, A.; Krastev, Z. Oral cavity and systemic diseases—Diabetes mellitus. *Biotechnol. Biotechnol. Equip.* **2011**, *25*, 2183–2186. [CrossRef]
3. Nathan, D.M. Long-term complications of diabetes mellitus. *N. Engl. J. Med.* **1993**, *328*, 1676–1685. [CrossRef] [PubMed]
4. Zimmet, P.; Alberti, K.G.; Magliano, D.J.; Bennett, P.H. Diabetes mellitus statistics on prevalence and mortality: Facts and fallacies. *Nat. Rev. Endocrinol.* **2016**, *12*, 616–622. [CrossRef] [PubMed]
5. Danaei, G.; Finucane, M.; Lu, Y.; Singh, G.; Cowan, M.; Paciorek, C.; Lin, J.; Farzadfar, F.; Khang, Y.H.; Stevens, G.; et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: Systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants. *Lancet* **2011**, *378*, 31–40. [CrossRef] [PubMed]
6. Iancu, I.; Mota, M.; Iancu, E. Method for the analysing of blood glucose dynamics in diabetes mellitus patients. In Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, Romania, 22–25 May 2008; Volume 3, pp. 60–65. [CrossRef]
7. Robertson, G.; Lehmann, E.D.; Sandham, W.; Hamilton, D. Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: A proof-of-concept pilot study. *J. Electr. Comput. Eng.* **2011**, *2011*, 681786. [CrossRef]
8. Vehi, J.; Mujahid, O.; Contreras, I. Artificial Intelligence and Machine Learning for Diabetes Decision Support. In *Advanced Bioscience and Biosystems for Detection and Management of Diabetes*; Sadasivuni, K.K., Cabibihan, J.J., Al-Ali, A.K.A.M., Malik, R.A., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 259–272. [CrossRef]
9. Alsahli, M.; Gerich, J.E. Hypoglycemia. *Endocrinol. Metab. Clin.* **2013**, *42*, 657–676. [CrossRef] [PubMed]
10. Yayan, E.H.; Zengin, M.; Karabulut, Y.E.; Akıncı, A. The relationship between the quality of life and depression levels of young people with type I diabetes. *Perspect. Psychiatr. Care* **2019**, *55*, 291–299. [CrossRef]
11. Shende, P.; Sahu, P.; Gaud, R. A technology roadmap of smart biosensors from conventional glucose monitoring systems. *Ther. Deliv.* **2017**, *8*, 411–423. [CrossRef]
12. Chen, C.; Zhao, X.L.; Li, Z.H.; Zhu, Z.G.; Qian, S.H.; Flewitt, A.J. Current and emerging technology for continuous glucose monitoring. *Sensors* **2017**, *17*, 182. [CrossRef]
13. Cappon, G.; Vettoretti, M.; Sparacino, G.; Facchinetti, A. Continuous glucose monitoring sensors for diabetes management: A review of technologies and applications. *Diabetes Metab. J.* **2019**, *43*, 383–397. [CrossRef] [PubMed]
14. Ibrahim, E.; Shouman, M.A.; Torkey, H.; El-Sayed, A. Handling missing and outliers values by enhanced algorithms for an accurate diabetic classification system. *Multimed. Tools Appl.* **2021**, *80*, 20125–20147. [CrossRef]
15. Saini, P.; Nagpal, B. Analysis of missing data and comparing the accuracy of imputation methods using wheat crop data. *Multimed. Tools Appl.* **2023**, *83*, 1–22. [CrossRef]
16. Little, R.J.; D'Agostino, R.; Cohen, M.L.; Dickersin, K.; Emerson, S.S.; Farrar, J.T.; Frangakis, C.; Hogan, J.W.; Molenberghs, G.; Murphy, S.A.; et al. The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* **2012**, *367*, 1355–1360. [CrossRef] [PubMed]
17. O'neill, R.; Temple, R. The prevention and treatment of missing data in clinical trials: An FDA perspective on the importance of dealing with it. *Clin. Pharmacol. Ther.* **2012**, *91*, 550–554. [CrossRef] [PubMed]
18. Noguer, J.; Contreras, I.; Mujahid, O.; Beneyto, A.; Vehi, J. Generation of Individualized Synthetic Data for Augmentation of the Type 1 Diabetes Data Sets Using Deep Learning Models. *Sensors* **2022**, *22*, 4944. [CrossRef] [PubMed]
19. An, Q.; Rahman, S.; Zhou, J.; Kang, J.J. A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors* **2023**, *23*, 4178. [CrossRef] [PubMed]
20. Modak, S.K.S.; Jha, V.K. Diabetes prediction model using machine learning techniques. *Multimed. Tools Appl.* **2023**, *83*, 1–27. [CrossRef]
21. Maniruzzaman, M.; Kumar, N.; Abedin, M.; Islam, M.; Suri, H.; El-Baz, A.; Suri, J. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Comput. Methods Programs Biomed.* **2017**, *152*, 23–34. [CrossRef]
22. Maniruzzaman, M.; Rahman, M.; Al-MehediHasan, M.; Suri, H.; Abedin, M.; El-Baz, A.; Suri, J. Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. *J. Med. Syst.* **2018**, *42*, 92. [CrossRef]
23. Srivastava, S.; Singh, S.; Suri, J. Healthcare text classification system and its performance evaluation: A source of better intelligence by characterizing healthcare text. *J. Med Syst.* **2018**, *42*, 97. [CrossRef] [PubMed]
24. Luo, G. Automatically explaining machine learning prediction results: A demonstration on type 2 diabetes risk prediction. *Health Inf. Sci. Syst.* **2016**, *4*, 2. [CrossRef] [PubMed]
25. Shakeel, P.; Baskar, S.; Dhulipala, V.; Jaber, M. Cloud based framework for diagnosis of diabetes mellitus using K-means clustering. *Health Inf. Sci. Syst.* **2018**, *6*, 16. [CrossRef] [PubMed]
26. Luo, G. MLBCD: A machine learning tool for big clinical data. *Health Inf. Sci. Syst.* **2015**, *3*, 3. [CrossRef] [PubMed]
27. Mujahid, O.; Contreras, I.; Vehi, J. Machine learning techniques for hypoglycemia prediction: Trends and challenges. *Sensors* **2021**, *21*, 546. [CrossRef] [PubMed]

28. Bertachi, A.; Viñals, C.; Biagi, L.; Contreras, I.; Vehí, J.; Conget, I.; Giménez, M. Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor. *Sensors* **2020**, *20*, 1705. [CrossRef] [PubMed]
29. Vahedi, M.R.; MacBride, K.B.; Wunsik, W.; Kim, Y.; Fong, C.; Padilla, A.J.; Pourhomayoun, M.; Zhong, A.; Kulkarni, S.; Arunachalam, S.; et al. Predicting glucose levels in patients with type1 diabetes based on physiological and activity data. In Proceedings of the 8th ACM MobiHoc 2018 Workshop on Pervasive Wireless Healthcare Workshop, Los Angeles, CA, USA, 25–28 June 2018; pp. 1–5. [CrossRef]
30. San, P.P.; Ling, S.H.; Nguyen, H.T. Deep learning framework for detection of hypoglycemic episodes in children with type 1 diabetes. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 3503–3506. [CrossRef]
31. Jin, Y.; Li, F.; Vimalananda, V.G.; Yu, H. Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: Empirical study. *JMIR Med. Inform.* **2019**, *7*, e14340. [CrossRef] [PubMed]
32. Quan, T.M.; Doike, T.; Bui, D.C.; Arata, S.; Kobayashi, A.; Islam, M.Z.; Niitsu, K. AI−based edge−intelligent hypoglycemia prediction system using alternate learning and inference method for blood glucose level data with low-periodicity. In Proceedings of the 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Hsinchu, Taiwan, 18–20 March 2019; pp. 201–206. [CrossRef]
33. Bertachi, A.; Biagi, L.; Contreras, I.; Luo, N.; Vehí, J. Prediction of Blood Glucose Levels And Nocturnal Hypoglycemia Using Physiological Models and Artificial Neural Networks. In Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data Co-Located with the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI 2018), Stockholm, Schweden, 13 July 2018; Bach, K., Bunescu, R.C., Farri, O., Guo, A., Hasan, S.A., Ibrahim, Z.M., Marling, C., Raffa, J., Rubin, J., Wu, H., Eds.; CEUR-WS.org; CEUR Workshop Proceedings, Sun SITE Central Europe, Germany. 2018; Volume 2148, pp. 85–90.
34. Oviedo, S.; Contreras, I.; Bertachi, A.; Quirós, C.; Giménez, M.; Conget, I.; Vehi, J. Minimizing postprandial hypoglycemia in Type 1 diabetes patients using multiple insulin injections and capillary blood glucose self-monitoring with machine learning techniques. *Comput. Methods Programs Biomed.* **2019**, *178*, 175–180. [CrossRef]
35. Vehí, J.; Contreras, I.; Oviedo, S.; Biagi, L.; Bertachi, A. Prediction and prevention of hypoglycaemic events in type-1 diabetic patients using machine learning. *Health Inform. J.* **2020**, *26*, 703–718. [CrossRef]
36. Mhaskar, H.N.; Pereverzyev, S.V.; Van der Walt, M.D. A deep learning approach to diabetic blood glucose prediction. *Front. Appl. Math. Stat.* **2017**, *3*, 14. [CrossRef]
37. Zhu, T.; Li, K.; Chen, J.; Herrero, P.; Georgiou, P. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. *J. Healthc. Inform. Res.* **2020**, *4*, 308–324. [CrossRef]
38. Mayo, M.; Chepulis, L.; Paul, R.G. Glycemic-aware metrics and oversampling techniques for predicting blood glucose levels using machine learning. *PLoS ONE* **2019**, *14*, e0225613. [CrossRef] [PubMed]
39. Li, K.; Liu, C.; Zhu, T.; Herrero, P.; Georgiou, P. GluNet: A deep learning framework for accurate glucose forecasting. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 414–423. [CrossRef]
40. Li, K.; Daniels, J.; Liu, C.; Herrero, P.; Georgiou, P. Convolutional recurrent neural networks for glucose prediction. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 603–613. [CrossRef]
41. Mosquera-Lopez, C.; Dodier, R.; Tyler, N.; Resalat, N.; Jacobs, P. Leveraging a big dataset to develop a recurrent neural network to predict adverse glycemic events in type 1 diabetes. *IEEE J. Biomed. Health Inform.* **2019**. [CrossRef]
42. Sisodia, D.; Sisodia, D.S. Prediction of diabetes using classification algorithms. *Procedia Comput. Sci.* **2018**, *132*, 1578–1585. [CrossRef]
43. Reddy, R.; Resalat, N.; Wilson, L.M.; Castle, J.R.; El Youssef, J.; Jacobs, P.G. Prediction of hypoglycemia during aerobic exercise in adults with type 1 diabetes. *J. Diabetes Sci. Technol.* **2019**, *13*, 919–927. [CrossRef] [PubMed]
44. Seo, W.; Lee, Y.B.; Lee, S.; Jin, S.M.; Park, S.M. A machine-learning approach to predict postprandial hypoglycemia. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–13. [CrossRef] [PubMed]
45. Jelinek, H.F.; Stranieri, A.; Yatsko, A.; Venkatraman, S. Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis. *Comput. Biol. Med.* **2016**, *75*, 90–97. [CrossRef]
46. Dave, D.; DeSalvo, D.J.; Haridas, B.; McKay, S.; Shenoy, A.; Koh, C.J.; Lawley, M.; Erraguntla, M. Feature-based machine learning model for real-time hypoglycemia prediction. *J. Diabetes Sci. Technol.* **2021**, *15*, 842–855. [CrossRef]
47. Chen, J.; Lalor, J.; Liu, W.; Druhl, E.; Granillo, E.; Vimalananda, V.G.; Yu, H. Detecting hypoglycemia incidents reported in patients' secure messages: Using cost-sensitive learning and oversampling to reduce data imbalance. *J. Med. Internet Res.* **2019**, *21*, e11990. [CrossRef] [PubMed]
48. Zhang, Y. Predicting occurrences of acute hypoglycemia during insulin therapy in the intensive care unit. In Proceedings of the 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 20–24 August 2008; pp. 3297–3300. [CrossRef]
49. Bashir, S.; Qamar, U.; Khan, F.H. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. *J. Biomed. Inform.* **2016**, *59*, 185–200. [CrossRef] [PubMed]
50. Lee, B.J.; Kim, J.Y. Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE J. Biomed. Health Inform.* **2015**, *20*, 39–46. [CrossRef] [PubMed]

51. Rau, H.H.; Hsu, C.Y.; Lin, Y.A.; Atique, S.; Fuad, A.; Wei, L.M.; Hsu, M.H. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. *Comput. Methods Programs Biomed.* **2016**, *125*, 58–65. [CrossRef] [PubMed]
52. Khan, F.A.; Zeb, K.; Al-Rakhami, M.; Derhab, A.; Bukhari, S.A.C. Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. *IEEE Access* **2021**, *9*, 43711–43735. [CrossRef]
53. Mosquera-Lopez, C.; Dodier, R.; Tyler, N.S.; Wilson, L.M.; El Youssef, J.; Castle, J.R.; Jacobs, P.G. Predicting and Preventing Nocturnal Hypoglycemia in Type 1 Diabetes Using Big Data Analytics and Decision Theoretic Analysis. *Diabetes Technol. Ther.* **2020**, *22*, 801–811. [CrossRef] [PubMed]
54. Wells, B.J.; Chagin, K.M.; Nowacki, A.S.; Kattan, M.W. Strategies for handling missing data in electronic health record derived data. *Egems* **2013**, *1*, 1035. [CrossRef] [PubMed]
55. Ma, S.; Schreiner, P.J.; Seaquist, E.R.; Ugurbil, M.; Zmora, R.; Chow, L.S. Multiple predictively equivalent risk models for handling missing data at time of prediction: With an application in severe hypoglycemia risk prediction for type 2 diabetes. *J. Biomed. Inform.* **2020**, *103*, 103379. [CrossRef] [PubMed]
56. Molenberghs, G.; Kenward, M. *Missing Data in Clinical Studies*; John Wiley & Sons: Hoboken, NJ, USA, 2007. [CrossRef]
57. Butt, H.; Khosa, I.; Iftikhar, M.A. Feature Transformation for Efficient Blood Glucose Prediction in Type 1 Diabetes Mellitus Patients. *Diagnostics* **2023**, *13*, 340. [CrossRef] [PubMed]
58. Estremera, E.; Cabrera, A.; Beneyto, A.; Vehi, J. A simulator with realistic and challenging scenarios for virtual T1D patients undergoing CSII and MDI therapy. *J. Biomed. Inform.* **2022**, *132*, 104141. [CrossRef]
59. Akturk, H.K.; Herrero, P.; Oliver, N.; Wise, H.; Eikermann, E.; Snell-Bergeon, J.; Shah, V.N. Impact of Different Types of Data Loss on Optimal Continuous Glucose Monitoring Sampling Duration. *Diabetes Technol. Ther.* **2022**, *24*, 749–753. [CrossRef] [PubMed]
60. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830. Available online: http://jmlr.org/papers/v12/pedregosa11a.html (accessed on 14 May 2024).
61. Chen, R.C.; Dewi, C.; Huang, S.W.; Caraka, R.E. Selecting critical features for data classification based on machine learning methods. *J. Big Data* **2020**, *7*, 1–26. [CrossRef]
62. Wee, B.F.; Sivakumar, S.; Lim, K.H.; Wong, W.; Juwono, F.H. Diabetes detection based on machine learning and deep learning approaches. *Multimed. Tools Appl.* **2023**, *83*, 1–33. [CrossRef]
63. Ergul Aydin, Z.; Kamisli Ozturk, Z. Filter-based feature selection methods in the presence of missing data for medical prediction models. *Multimed. Tools Appl.* **2024**, *83*, 24187–24216. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.