applied sciences

MDPI

Article

Enhancing Nocturnal Hypoglycemia Prediction Through Ensemble Feature Selection and Synthetic Data Augmentation

Bryan Peralta-Arboleda [1], David Chushig-Muzo [1,*], Francisco J. Lara-Abelenda [1], Carmelo Betancort Acosta [2,3], Ana M. Wägner [2,3], Conceição Granja [4] and Cristina Soguero-Ruiz [1,*]

[1] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, 28943 Madrid, Spain; bryan.peralta@urjc.es (B.P.-A.); francisco.lara@urjc.es (F.J.L.-A.)
[2] Instituto Universitario de Investigaciones Biomédicas y Sanitarias, Universidad de Las Palmas de Gran Canaria, 35016 Las Palmas de Gran Canaria, Spain; carmelobetancort@gmail.com (C.B.A.); ana.wagner@ulpgc.es (A.M.W.)
[3] Endocrinología y Nutrición, Complejo Hospitalario Universitario Insular Materno Infantil, 35016 Las Palmas de Gran Canaria, Spain
[4] Norwegian Centre for E-Health Research, University Hospital of North Norway, 9019 Tromso, Norway; conceicao.granja@ehealthresearch.no
* Correspondence: david.chushig@urjc.es (D.C.-M.); cristina.soguero@urjc.es (C.S.-R.)

Abstract

Nocturnal hypoglycemia (NH), which occurs when blood glucose levels fall below 70 mg/dL during sleep, remains a common complication in people with type 1 diabetes. Despite advances in continuous glucose monitoring (CGM), controlling both glucose and glycemic events remains challenging. Artificial Intelligence (AI) models are promising for identifying CGM patterns associated with NH and support its prediction, thereby reducing associated complications. However, two main challenges limit the performance of AI-based models for NH: (i) the class imbalance caused by the low number of NH events in the datasets; and (ii) the high-dimensionality caused by the extraction of CGM-derived features. To address this, we propose a two-stage methodology that combines ensemble feature selection (FS) methods and synthetic data augmentation techniques. We employed CGM features extracted using two window-based approaches (non-overlapping and overlapping) with three, four, and six windows. Model interpretability is provided through feature analysis using Shapley additive explanations (SHAP), enabling the identification of associated NH factors. We employed CGM data from 52 individuals with T1D, collected at Complejo Hospitalario Insular-Materno Infantil de Las Palmas de Gran Canaria, Spain. Among the evaluated configurations, the six-window configuration yielded the highest reported predictive performance. The two LR models examined in the SHAP analysis, both trained with RUS and overlapping windows, achieved test-set AUCROC values of 0.7403 using all features and 0.7400 using the selected feature subset, calculated from continuous prediction scores. Our findings extend prior evidence on temporal CGM representations to an adult NH setting not restricted to exercise-associated nights, using bootstrap-based Relief FS and alternative windowing configurations. Our interpretable methodology provides a clinically relevant approach for the early prediction of NH, supporting decision-making and enhancing the effectiveness of CGM-based monitoring systems. The proposed methodology may support timely preventive interventions and contribute to improving patient safety and confidence in diabetes management.

Keywords: nocturnal hypoglycemia; type 1 diabetes; continuous glucose monitoring; ensemble feature selection; data augmentation; model interpretability; SHAP; TabPFN

Check for updates

Academic Editor: Grazia Maugeri

Received: 14 August 2026
Revised: 23 October 2026
Accepted: 30 October 2026
Published: 5 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Type 1 diabetes (T1D) is an autoimmune disease characterized by the destruction of pancreatic beta cells, resulting in absolute/partial insulin deficiency [1]. Hypoglycemia remains one of the most prevalent complications in T1D management, defined as blood glucose below 70 mg/dL for at least 15 min. Nocturnal hypoglycemia (NH) is especially concerning because it occurs during sleep, when physiological counterregulatory responses are diminished, and symptoms may remain unnoticed until an episode occurs [2]. NH is associated with impaired sleep quality, increased fear of hypoglycemia, and reduced quality of life for patients [3].

Continuous glucose monitoring (CGM) systems represent a significant advancement in diabetes management, providing interstitial glucose measurements every 5–15 min to support glycemic control [4]. As key parts of the healthcare Internet of Things, CGM devices continuously generate and transmit physiological data, providing the foundation for data-driven approaches. Artificial Intelligence (AI) has created valuable opportunities to support clinical decision-making. For instance, several studies have employed AI-based models and CGM data for glucose forecasting [5], severe hypoglycemia detection [6], characterizing the impact of physical activity on glucose [7], detecting glycemic excursions [8], and supporting insulin dosing decisions [9]. Despite these advances, NH prediction remains an active and open area of research with opportunities for further improvement.

Most NH prediction studies have used CGM-derived features to capture glycemic dynamics. Established metrics, including time in range, time below range, time above range, and glycemic variability metrics (e.g., standard deviation, coefficient of variation), have been used for predicting glycemic events [10,11]. However, when these features are computed across a range of temporal windows, this process results in high-dimensionality, thereby posing challenges for training models. Additionally, although CGM systems offer valuable data, their efficacy for training predictive models is often hampered by the scarcity of NH events, which introduces a severe class imbalance problem (CIP) in supervised learning. Therefore, the development of NH prediction models remains a challenging task, mainly hindered by two limitations: (i) the high-dimensionality of CGM-derived features; and (ii) severe class imbalance.

Beyond predictive performance, Explainable Artificial Intelligence (XAI) improves the interpretability of Machine Learning (ML) models by identifying features that drive risk predictions [12]. Shapley additive explanations (SHAP) provides local and global feature attributions and has been applied to CGM-based hypoglycemia prediction in T1D [12,13]. More recently, SHAP has also been applied to NH prediction in older adults with Type 2 diabetes, identifying clinically relevant features such as the lowest daytime glucose level, fasting glucose, and previous daytime hypoglycemia [14]. These findings highlight the potential of SHAP to link model predictions with interpretable glycemic risk factors [12].

This study proposes a two-stage methodology that combines ensemble feature selection (FS), window-based feature extraction, and synthetic data augmentation techniques for NH prediction from CGM data. In this paper, a homogeneous ensemble applies Relief to bootstrap samples of the training data and combines the resulting selections through voting. The study evaluates this FS procedure across temporal representations, classifiers, and resampling configurations for NH prediction. We compared two window-based feature representations: (i) non-overlapping windows, which divide the input sequence into separate consecutive segments; and (ii) nested overlapping windows, which share the final sequence position and begin at progressively later positions. We address the CIP inherent to NH prediction through synthetic data augmentation techniques, including Random Undersampling (RUS) and the Synthetic Minority Over-sampling Technique (SMOTE). The

study used CGM data from 52 individuals with T1D enrolled at the Complejo Hospitalario Insular-Materno Infantil de Las Palmas de Gran Canaria, Spain.

Temporally structured CGM features, FS, class balancing, and SHAP have previously been combined for post-exercise NH prediction in a pediatric cohort by Piersanti et al. [15]. Building on this precedent, the present study evaluates bootstrap-based Relief voting for NH prediction in an adult cohort, without restricting the analysis to exercise-associated nights, and compares performance across temporal representations and target imbalance ratios (IRs).

The main contributions of this study are summarized as follows:

- We evaluate bootstrap-based Relief FS with voting aggregation for NH prediction in adults, comparing the selected feature subset with the full set of extracted features under RUS and SMOTE, across different target IRs.
- We extend the evaluation of temporal CGM representations to the adult cohort studied by comparing windowing configurations with and without overlap.
- We employ an XAI method, SHAP, to characterize the contributions of CGM-derived features to model predictions under the evaluated temporal configurations.
- We evaluate the proposed methodology on a real-world CGM dataset comprising 52 adults with T1D and 8875 participant–day samples, assessing its predictive performance for NH under different experimental configurations.

The remainder of the paper is organized as follows. Section 2 reviews related work on NH prediction. Section 3 describes the dataset and feature extraction process. Section 4 presents the proposed methodology, and Section 5 presents the experimental results. Finally, the discussion and conclusions are presented in Sections 6 and 7, respectively.

2. Related Work

This section reviews recent state-of-the-art studies on NH prediction using CGM data. Table 1 summarizes these studies, highlighting the datasets used, prediction horizons, predictive models, and the adoption of FS and data augmentation techniques.

Table 1. A summary of studies focused on predicting nocturnal hypoglycemia. These works were sorted by publication date. The Samples/Observations column reports the number of nights, CGM observations, or evaluation instances as reported in each study; differences in the unit of analysis are indicated where necessary.

Study	Year	Individuals	Samples/ Observations	Prediction Horizon	Models/ Methods	Feature Selection	Data Augmentation
Vu et al. [16]	2019	10,000	1,000,000	6 h	RF	⊖	◦
Mosquera-Lopez et al. [17]	2020	134	22,804	6 h	SVR	⊖	◦
Jensen et al. [18]	2020	463	4271	6 h	LDA	⊖	◦
Berikov et al. [19]	2022	406	6451	15/30 min	RF, LASSO, MLP	•	◦
Mosquera-Lopez et al. [20]	2024	366	44,154	8 h	SVR, ENN	⊖	◦
Kozinetz et al. [21]	2024	380	380	30 min	CNNs	⊖	◦
Yu et al. [22]	2024	12	7000	30 min	Dynamic Attention DNN	⊖	◦
Leutheuser et al. [23]	2024	13	66	8 h	RF, LR, MLP, RNN	⊖	◦

Table 1. *Cont.*

Study	Year	Individuals	Samples/ Observations	Prediction Horizon	Models/ Methods	Feature Selection	Data Augmentation
Voegeli et al. [24]	2025	12	–	8 h	RF, LSTM, CNN, DailyLSTM DailyCNN	◦	•
Afentakis et al. [25]	2025	37	222,000	–	RF, SVM	•	◦
Lara et al. [26]	2025	52	8875	6 h	TSC	⊖	◦
Tziavaras et al. [27]	2026	12	193,536	6 h	LSTM + AE	⊖	◦
Piersanti et al. [15]	2026	55	49	8 h	RF	•	•
Roquemen-Echeverri et al. [28]	2026	20	502	8 h	NN	◦	◦
Herrero et al. [29]	2026	204	29,929 *	7 h	Night Low Predict + Digital Twin	◦	•

Source: Adapted and expanded from a table in Lara-Abelenda et al. [26], with additional studies and columns indicating whether each study used feature selection and data augmentation. **Symbols:** ◦ = Not covered; ⊖ = Partially covered; • = Fully covered; – it is not reported in the study. **Acronyms:** Random Forest (RF); Support Vector Regression (SVR); Linear Discriminant Analysis (LDA); Least Absolute Shrinkage and Selection Operator (LASSO); Multi-Layer Perceptron (MLP); Support Vector Machine (SVM); evidential neural network (ENN); Neural Network (NN); Convolutional Neural Network (CNN); Deep Neural Network (DNN); logistic regression (LR); Recurrent Neural Network (RNN); Long Short-Term Memory (LSTM); Autoencoder (AE); Time Series Classifiers (TSCs). **Notes:** * Herrero et al. [29] report 29,929 simulated days from 204 digital twins rather than independent observed CGM nights.

Most studies in NH prediction are based on CGM-derived features combined with traditional ML models. For instance, Vu et al. [16] trained a Random Forest (RF) on more than one million nights of data, reporting the Area Under the Receiver Operating Characteristic Curve (AUCROC) values of 0.90 (during the early night) and 0.75 (during the late night). Jensen et al. [18] and Mosquera-Lopez et al. [17] also investigated CGM-based NH prediction, reporting AUCROC values of 0.79 and 0.86, respectively.

Other studies have explored shorter prediction horizons and embedded FS. For example, Berikov et al. [19] evaluated RF, LASSO, and MLP, using LASSO as an embedded FS method, and reported an AUCROC value of 0.97 for 15 min horizons. Mosquera et al. [20] trained an evidential neural network, achieving AUCROC values of 0.71 and 0.80 across different phases of the night. More recent studies incorporate deep learning architectures, including convolutional networks [24], recurrent networks [27], and attention-based models [22]. Lara et al. [26] evaluated a range of time series classification models for NH prediction, reporting a maximum AUCROC of 0.703. Among the evaluated methods, interval-based and feature-based approaches achieved the best predictive performance.

First, FS has been investigated to reduce the dimensionality of CGM-derived representations in NH prediction [15,25]. Second, class imbalance is a fundamental characteristic of NH prediction, as hypoglycemic events represent a small proportion of nocturnal observations. Resampling has also been incorporated into NH prediction pipelines, including SMOTE-based class balancing combined with feature selection in Piersanti et al. [15].

More recent studies have focused on making NH prediction more clinically useful. Piersanti et al. [15] developed an RF-based approach using CGM-derived biomarkers extracted from exercise and recovery windows to predict NH within an 8 h post-exercise

horizon. Roquemen-Echeverri et al. [28] applied SHAP to interpret predictions from a previously developed evidential neural network, identifying bedtime glucose, hypo- and hyperglycemia exposure, glycemic variability, and physical activity as relevant features. Herrero et al. [29] evaluated the clinical usefulness of the Night Low Predict algorithm through a digital-twin simulation framework. In this setting, virtual representations of individuals with T1D were used to estimate the potential glycemic effects of acting on 7 h NH risk predictions. Collectively, these studies addressed complementary aspects of NH prediction: feature-based modeling [15], model interpretability [28], and simulation-based evaluation of preventive actions [29].

Several limitations remain in this body of work. First, reported AUCROC values are not directly comparable, since studies differ in prediction horizon (from 15 min to 8 h), unit of analysis (nights, CGM observations, or simulated days), and cohort size, with four studies including only 12 or 13 individuals [22–24,27]. Second, FS and class-imbalance handling have mostly been addressed separately. Only three of the reviewed studies fully incorporated FS [15,19,25], and data augmentation was fully addressed in three [15,24,29]. Piersanti et al. [15] is the only study combining both, and it was restricted to post-exercise nights in a pediatric cohort. Third, the stability of selected features under data perturbation and its interaction with resampling and temporal representation have not been examined. The present study addresses these gaps in an adult cohort with NH not restricted to exercise-associated nights. It evaluates bootstrap-based Relief voting jointly with RUS and SMOTE across target IRs and under overlapping and non-overlapping windows, using SHAP to characterize the resulting feature contributions.

3. Dataset Description and Preprocessing

3.1. Dataset

We employed CGM data from 52 individuals with T1D recruited at the Complejo Hospitalario Insular-Materno Infantil de Las Palmas de Gran Canaria. All participants were recorded as adults in the study records. Participants were invited during scheduled diabetes clinic visits by physicians participating in the WARIFA project [30], and written informed consent was obtained from all individuals. The Provincial Ethics Committee of Las Palmas determined that the study was exempt from ethical review, as it did not fall under the scope of Biomedical Research regulations. This project, funded by the European Union, aimed to build personalized early risk prediction models that provide personalized recommendations for the prevention of chronic diseases. CGM data from enrolled participants were securely stored on a server accessible only via a virtual private network, with access limited to WARIFA personnel.

The daytime period was defined as the interval from 06:01 to 23:59 [26]. Within this period, the CGM observations used for NH prediction spanned from 12:00 to 23:45, covering a nominal 12 h observation window sampled at 15 min intervals. The subsequent nocturnal period used to define the NH label extended from 00:00 to 06:00. The predictor and nocturnal outcome intervals were defined using fixed clock times for all participants, rather than individual sleep onset or awakening times.

Following established clinical criteria, NH was defined as the occurrence of at least three consecutive CGM readings below 70 mg/dL during the nocturnal period. This minimizes false positives from CGM devices, ensuring that labeled NH events reflect clinically relevant hypoglycemia. A total of 8875 CGM records were included, each corresponding to a participant–day sample represented by 48 preprocessed CGM values and an associated NH label. Of these samples, 6990 (78.8%) were labeled as non-NH and 1885 (21.2%) as NH, reflecting the class imbalance. Table 2 shows the total number and proportion of samples labeled as non-NH and NH.

Table 2. Summary of the dataset used for NH prediction. Percentages are calculated over the total number of participant–day samples.

Label	Number	Percentage (%)
Non-hypoglycemia (No-NH)	6990	78.8
Nocturnal Hypoglycemia (NH)	1885	21.2
	8875	100.0

3.2. Temporal Window Approach and Extraction of CGM-Derived Features

Each participant–day sample comprised 48 previously prepared CGM values, corresponding to a 12 h observation window within the daytime period (06:01–23:59) [26]. For feature extraction, these values were assigned computational timestamps at 15 min intervals, from 12:00 to 23:45. We evaluated configurations with three, four, or six windows. Windows were defined either as consecutive non-overlapping segments or as overlapping nested suffixes sharing a common final endpoint (Figure 1).

1. Non-overlapping windows: features were computed separately within consecutive, disjoint intervals.
2. Overlapping windows: Successive windows started at progressively later boundaries and shared the same final timestamp, resulting in nested windows of decreasing duration. Each shorter window was fully contained within the preceding longer window and therefore represented a progressively more recent portion of the CGM history.

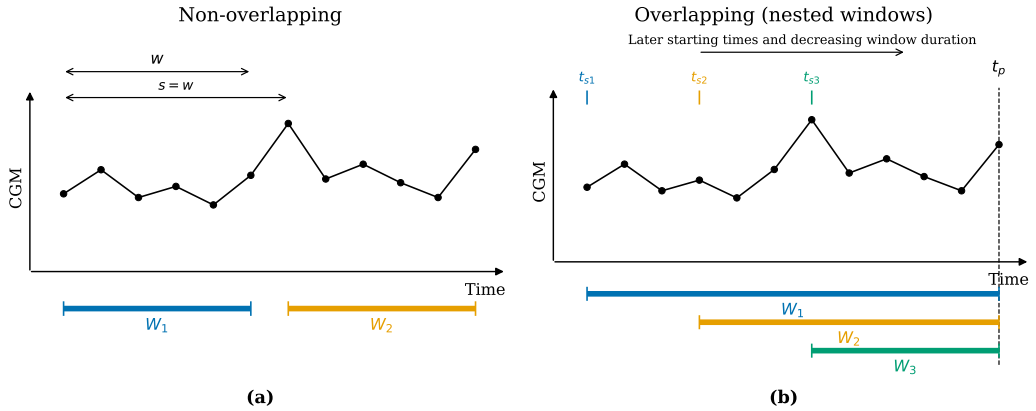


Figure 1. Schematic of temporal window-based feature extraction from continuous glucose monitoring (CGM) data. (**a**) In the non-overlapping approach, consecutive windows represent disjoint segments of the CGM sequence without temporal overlap. (**b**) In the overlapping approach, windows share the prediction time t_p as a common endpoint, while their starting times are progressively shifted forward, resulting in windows of decreasing duration. Each shorter window is fully contained within the preceding longer window and captures a progressively more recent portion of the CGM history. Features are computed separately for each temporal window to construct the model input representation.

Windows were indexed using zero-based sequence positions according to increasing start position. In the six-window configuration, the non-overlapping windows comprised the inclusive positions 0–7, 8–15, 16–23, 24–31, 32–39, and 40–47. The corresponding nested windows comprised positions 0–47, 8–47, 16–47, 24–47, 32–47, and 40–47. Features were computed separately for each window and then concatenated into a single feature vector for each sample. Time-dependent features were calculated using the assigned computational timestamps.

To capture glucose dynamics in different time windows, we used GlucoStats (version 1.0.0) [31], an open-source Python library developed by members of our research group. The experiments were performed using Python (version 3.10.16). Its support for window-

based feature extraction and compatibility with scikit-learn motivated its integration into our analysis pipeline. For each temporal window, CGM-derived features were computed. The metrics summarized in Table 3 are grouped into five categories: risk indices, glycemic control metrics, range-based metrics, glycemic variability metrics, and descriptive and signal statistics. Features with more than 60% of zero values were excluded. Missing values in continuous features were imputed using the mean (only using training data).

Several software packages are available for calculating CGM-derived metrics, including iglu [32], cgmanalysis [33], and cgmquantify [34], which were examined in the comparative review by Piersanti et al. [34]. That review showed that CGM-derived metrics may vary across software implementations due to differences in algorithms, parameter settings, preprocessing procedures, and units. These discrepancies may particularly affect measures such as Mean Amplitude of Glycemic Excursions (MAGE) and Area Under the Glucose Curve (AUC) [34]. GlucoStats was not included among the tools evaluated in that comparative review. Although its original publication demonstrated its functionality using a public CGM dataset [31], this evidence does not establish numerical equivalence with the software packages evaluated by Piersanti et al. No direct cross-package comparison was conducted in the present study.

Table 3. Summary of CGM-derived metrics present in the stored feature matrices before zero-value filtering and ensemble feature selection. Availability varies across temporal configurations.

Feature	Description
Risk indices	
HBGI	High blood glucose index
LBGI	Low blood glucose index
max_hbgi	Maximum observation-level high-glucose risk within the window
max_lbgi	Maximum observation-level low-glucose risk within the window
BGRI	Blood glucose risk index
GRADE	Glycemic risk assessment in diabetes equation
GRADE_hyper	Percentage of total GRADE attributable to hyperglycemia
GRADE_hypo	Percentage of total GRADE attributable to hypoglycemia
GRADE_eu	Percentage of total GRADE attributable to euglycemia
Glycemic control metrics	
GMI	Glucose management indicator
eA1c	Estimated HbA1c
IGC	Index of glycemic control
hyper_index	Hyperglycemia index
hypo_index	Hypoglycemia index
m_value	M-value, combining deviations from a reference glucose level with the glucose range
j_index	J-index, combining mean glucose and its standard deviation
Range-based metrics	
T_IR	Time in range
T_BR	Time below range
T_AR	Time above range
T_OR	Time out of range
PT_IR	Percentage of time in range
PT_BR	Percentage of time below range
PT_AR	Percentage of time above range
PT_OR	Percentage of time out of range
N_IR	Number of observations in range
N_BR	Number of observations below range
N_AR	Number of observations above range
N_OR	Number of observations out of range
PN_IR	Percentage of observations in range
PN_BR	Percentage of observations below range
PN_AR	Percentage of observations above range
PN_OR	Percentage of observations out of range

Table 3. *Cont.*

Feature	Description
Glycemic variability metrics	
CV	Coefficient of variation
ef	Number of significant glucose excursions within the window
GVP	Glycemic variability percentage
DT	Sum of absolute differences between consecutive glucose values
MAG	Mean absolute glucose change per unit time (mg/dL/h)
Descriptive and signal statistics	
mean	Mean glucose value
SD	Standard deviation of glucose values
DFA	Detrended fluctuation analysis scaling exponent
AUC	Area under the glucose curve
min	Minimum glucose value
max	Maximum glucose value
max_diff	Difference between the maximum and minimum glucose values within the window
quartile_0.25	First quartile (25th percentile)
quartile_0.5	Median (50th percentile)
quartile_0.75	Third quartile (75th percentile)
IQR	Interquartile range

The study specified glucose thresholds of 70 and 180 mg/dL. Prefixes T_, PT_, N_, and PN_ denote elapsed time, percentage of time, number of observations, and percentage of observations, respectively. DFA was present only in some overlapping-window configurations.

4. Methods

This section presents the methods used in this study. Figure 2 illustrates the proposed workflow, which integrates window-based feature extraction from CGM data, data resampling to address class imbalance, supervised classification, and ensemble FS.

4.1. Resampling Techniques

Formally, let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the dataset, where N is the number of samples, $\mathbf{x}_i$ denotes the vector of CGM-derived features for the i-th sample, and $y_i \in \{0, 1\}$ is the class label. We define $y = 1$ as the occurrence of NH (minority class) and $y = 0$ as non-NH (majority class). The dataset is partitioned into a training set $\mathcal{D}_{train}$ and a test set $\mathcal{D}_{test}$. For the binary classification task, let N_{min} and N_{maj} denote the numbers of minority- and majority-class samples in $\mathcal{D}_{\text{train}}$, respectively (where $N_{\text{min}} < N_{\text{maj}}$). The IR is defined as follows:

$$\text{IR} = \frac{N_{\text{min}}}{N_{\text{maj}}}. \tag{1}$$

Throughout this study, IR is used to quantify the degree of class imbalance in the training data. To evaluate the effect of different imbalance levels, a resampling method $\mathcal{R}$ is applied exclusively to the training set to obtain

$$\mathcal{D}'_{\text{train}} = \mathcal{R}\left(\mathcal{D}_{\text{train}}; \text{IR}_{\text{target}}\right), \tag{2}$$

where $\text{IR}_{\text{target}}$ denotes the desired imbalance ratio in the resampled training set. Depending on the selected resampling technique, $\mathcal{R}(\cdot)$ modifies the class distribution by generating synthetic minority-class samples or removing majority-class samples. Resampling is performed only on $\mathcal{D}_{\text{train}}$, while the test set remains unchanged to preserve the original class distribution during performance evaluation.

In this study, two resampling techniques were employed: RUS and SMOTE. These established techniques were included to compare alternative class imbalance handling strategies across target IRs, rather than as novel resampling methods. RUS addresses class imbalance by randomly removing samples from the majority class (No-NH events), thereby reducing its dominance in the training data and producing a less imbalanced class distribution.

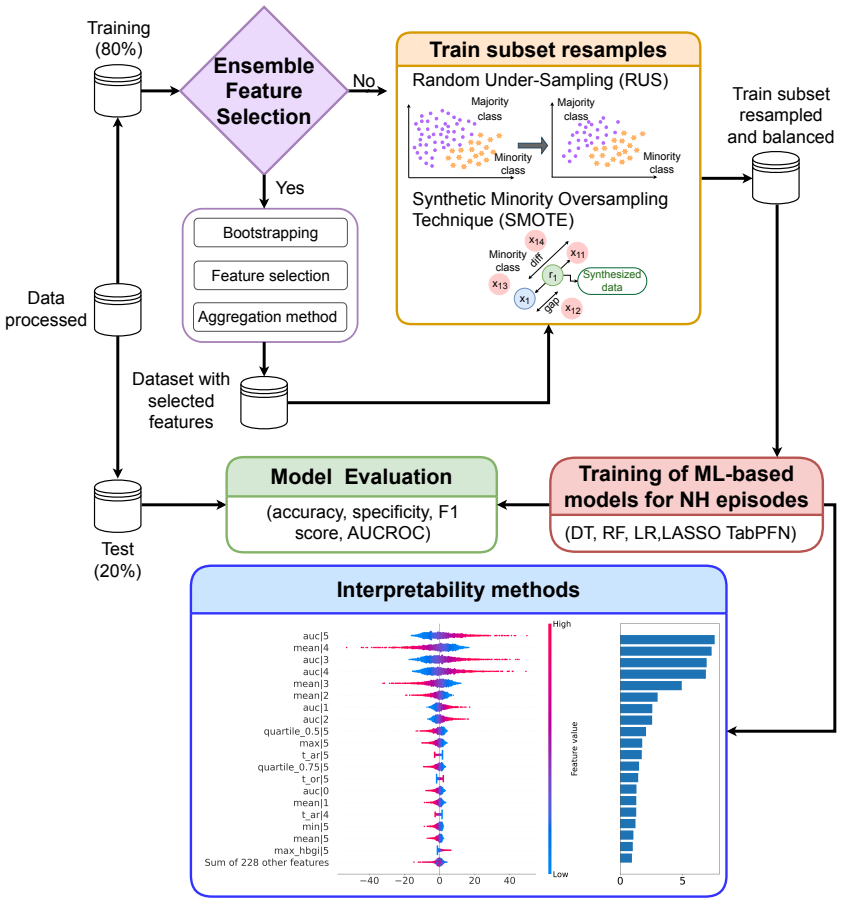


Figure 2. Overview of the workflow for nocturnal hypoglycemia (NH) prediction from continuous glucose monitoring (CGM)-derived features. Participant–day samples are partitioned into training and test subsets, with individual participants potentially represented in both. In the ensemble feature selection (FS) experiments, training-based standardization and bootstrap-based Relief selection precede random undersampling (RUS) or synthetic minority oversampling (SMOTE). Resampling is restricted to the training data. Classifiers are evaluated on the held-out test data, and Shapley additive explanations (SHAP) is used for post hoc interpretation.

SMOTE was used to generate synthetic minority-class samples (NH events). SMOTE is an interpolation method that consists of three key stages [35]: nearest neighbor identification, interpolation vector computation, and synthetic sample creation. For each minority class sample $\mathbf{x}_i$, the algorithm first identifies its k nearest neighbors $\{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ik}\}$ within the same class using the Euclidean distance [35]. New samples are then generated by linearly interpolating each instance with its neighbors [36]:

$$\tilde{\mathbf{x}} = \mathbf{x}_i + \lambda \left(\mathbf{x}_{ij} - \mathbf{x}_i\right), \quad \lambda \sim \mathcal{U}(0,1), \tag{3}$$

where λ represents a uniformly distributed interpolation coefficient. This linear interpolation creates synthetic instances along the line segments connecting existing minority class samples in the feature space. In the ensemble FS experiments, SMOTE operated on the selected standardized training features, after scaling and FS. The test set was not resampled. Two main hyperparameters are used for the algorithm: (i) k, the number of nearest neighbors; and (ii) the IR to determine the number of synthetic samples to be generated.

By applying RUS and SMOTE with different target IRs, the class distribution of the training set is systematically controlled while preserving the original test distribution. This experimental design enables a comprehensive evaluation of the impact of class imbalance on classifier performance, facilitates the assessment of model robustness across

different imbalance levels, and supports the identification of the IR that yields the best predictive performance.

4.2. Feature Selection for Addressing High-Dimensional CGM-Derived Features

For the ensemble FS experiments, z-score standardization parameters for CGM-derived features were estimated exclusively from the training set. Specifically, each feature was centered by subtracting the training-set mean and scaled by the corresponding standard deviation. Relief-based feature ranking was performed exclusively on the standardized training data, while the test set was transformed using the same training-derived scaling parameters, preventing test-set information from influencing preprocessing and FS [37]. In experiments without ensemble FS, explicit standardization was applied only to LR and LASSO after resampling, whereas DT, RF, and TabPFN did not receive additional external scaling. In the ensemble FS pipeline, LR and LASSO underwent an additional standardization step after resampling.

To reduce dimensionality and enable a clinical interpretation, we employed a homogeneous ensemble FS where multiple bootstrap subsets are generated from the training dataset, and an FS method is then applied to each subset [38]. We adopted a homogeneous ensemble in which Relief was applied to multiple bootstrap samples of the training data. The resulting feature selections were combined through voting to obtain a consensus ranking, as detailed below [39]. Figure 3 provides a schematic of homogeneous ensemble FS.

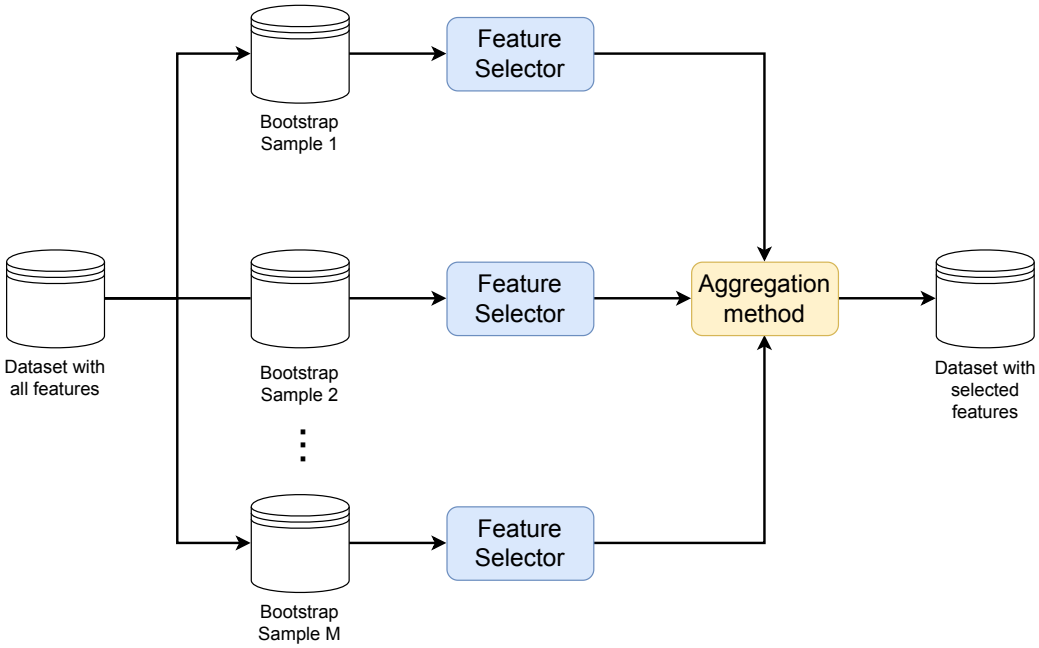


Figure 3. Homogeneous ensemble feature selection (FS) using Relief and bootstrap voting. Multiple bootstrap samples are drawn with replacement from the training data, and Relief ranks the features within each sample. Votes indicate inclusion among the top-ranked features of each bootstrap sample. Aggregated votes determine a consensus ranking from which the retained feature subset is selected.

Formally, let $\mathcal{D}_{train}^{(b)}$ denote the b-th bootstrap sample drawn with replacement from the training set $\mathcal{D}_{train}$, where $b = 1, \ldots, B$ and B is the total number of bootstrap subsets. For each bootstrap subset, feature importance was estimated using Relief, which evaluates features through differences between neighboring instances of the same and different classes [40]. Relief assessed local class discrimination, whereas bootstrap voting prioritized features that were consistently selected across perturbations of the training data. In contrast to the Variance Inflation Factor–Minimum Redundancy Maximum Relevance (VIF-mRMR) pipeline

used by Piersanti et al. [15], which addresses multicollinearity and feature redundancy, our approach produces a consensus ranking for feature-subset experiments. It does not imply explicit redundancy removal or superiority over VIF–mRMR.

Applying Relief to $\mathcal{D}_{train}^{(b)}$ produces a ranking of the D_f features according to their relevance scores. Let $\mathcal{T}^{(b)}$ denote the subset containing the top-K-ranked features obtained from the b-th bootstrap sample. The ensemble FS approach aggregates the individual rankings through a voting strategy. Specifically, a binary vote variable is assigned to each feature as

$$z_j^{(b)} = \begin{cases} 1, & \text{if } j \in \mathcal{T}^{(b)}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $j = 1, \ldots, D_f$ indexes the features. The total number of votes received by feature j across all bootstrap samples is then computed as

$$v_j = \sum_{b=1}^{B} z_j^{(b)}, \tag{5}$$

where $v_j \in 0, \ldots, B$. Features receiving larger values of v_j are selected more consistently across the bootstrap samples and are therefore considered more stable and relevant for the prediction task. Finally, the features are sorted in descending order of v_j to produce a consensus ranking of the CGM-derived features, which serves as the basis for constructing the reduced feature subsets used in the NH prediction task.

4.3. Model Interpretability Through Shapley Additive Explanations

To improve model interpretability, we employed SHAP, a post hoc explainability method based on cooperative game theory that quantifies the contribution of each feature to predictions [13]. Let $f(\mathbf{x})$ denote a trained classifier operating on an input feature vector $\mathbf{x} \in \mathbb{R}^D$, where D denotes the dimensionality of the input space. For a given sample $\mathbf{x}^{(i)}$, SHAP explains the model output by decomposing it into the additive contributions of the input features [41]:

$$f(\mathbf{x}^{(i)}) = \phi_0 + \sum_{j=1}^{D} \phi_j^{(i)}, \tag{6}$$

where $\phi_0 = \mathbb{E}[f(\mathbf{x})]$ is the baseline prediction, corresponding to the expected model output over a dataset, and $\phi_j^{(i)} \in \mathbb{R}$ is the SHAP value associated with feature j for sample $\mathbf{x}^{(i)}$. Each SHAP value quantifies the contribution of feature j to the deviation of the prediction from the baseline value. When the classifier predicts the probability of NH, the sign of the SHAP value indicates the direction of the feature's influence: positive values increase the predicted probability of NH, whereas negative values decrease it.

5. Results

In this section, we present the experimental setup as well as the predictive results using ensemble FS, window-based approaches and resampling techniques. Finally, we present an interpretability analysis using SHAP and CGM-derived features.

5.1. Experimental Setup

The dataset was partitioned into training and test subsets containing 80% and 20% of the participant–day samples, respectively, using stratification by outcome. Each participant–day sample was assigned exclusively to one subset. The split was performed at the participant–day level rather than the participant level; consequently, different days from the same participant could appear in both training and test sets. To address class imbalance, two resampling approaches were considered: RUS and SMOTE. Resampling was applied

only to the training subset. The following models were employed: Decision Tree (DT), logistic regression (LR), RF, LASSO, and Tabular Prior-Data Fitted Network (TabPFN). The hyperparameter search space considered for each model is summarized in Table 4. Model selection was conducted using three-fold cross-validation, optimizing the AUCROC. To address the class imbalance problem, we evaluated IRs in the range of 0.6–1.0 to investigate the effect of different class distributions on model performance.

Table 4. Hyperparameter configurations evaluated for the classifiers used in nocturnal hypoglycemia prediction.

Classifier	Hyperparameter	Values
DT	Maximum depth	[2, 4, 6, 8, 10, 12]
	Minimum samples leaf	[1, 2, 4, 6, 8]
	Split criterion	[gini, entropy]
LR	Regularization penalty	[l1, l2]
	Inverse regularization parameter (C)	$[1 \times 10^{-4}, 3.59 \times 10^{-4}, 1.29 \times 10^{-3}, 4.64 \times 10^{-3}, 1.66 \times 10^{-2}, 5.99 \times 10^{-2}, 0.215, 0.774, 2.78, 10]$
LASSO	Regularization strength (alpha)	[0.0001, 0.01, 1, 5, 10, 20]
RF	Number of estimators	[40, 60, 80, 100, 150, 200]
	Maximum depth	[2, 4, 6, 8, 10]
	Split criterion	[gini, entropy]
	Minimum samples split	[2, 4, 6, 8, 16]
TabPFN	Number of estimators	[40, 50, 100]

The predictive performance was measured using recall, specificity and the AUCROC, which are defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{TN + FP},$$

$$\text{AUCROC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}$$

where TP, FP, TN, and FN indicate the number of true positives, false positives, true negatives, and false negatives, respectively. For the ROC curve, the True Positive Rate (TPR) and the False Positive Rate (FPR) were computed as follows: TPR = Recall and FPR = 1 − Specificity.

5.2. Classification Results with Different Window Configurations

Figure 4 presents the NH prediction results obtained using overlapping and non-overlapping windows configurations with three, four, and six windows. Regarding AUCROC (see Figure 4a,b), LR and LASSO achieved the highest mean performance across all window configurations, with values ranging from 0.65 to 0.68. RF and TabPFN followed closely, yielding AUCROC values between 0.62 and 0.66, whereas DT consistently exhibited the lowest performance, with mean AUCROC values ranging from 0.60 to 0.63.

For the non-overlapping windows (see Figure 4a), most classifiers exhibited a gradual increase in AUCROC as the number of windows increased from three to six, with the highest values consistently achieved using the six-window configuration. A similar trend was observed for the overlapping windows (see Figure 4b), where the six-window configuration again produced the highest (or near-highest) AUCROC values across classifiers. In contrast,

DT exhibited greater variability across random seeds, particularly for configurations with fewer windows, indicating less stable performance.

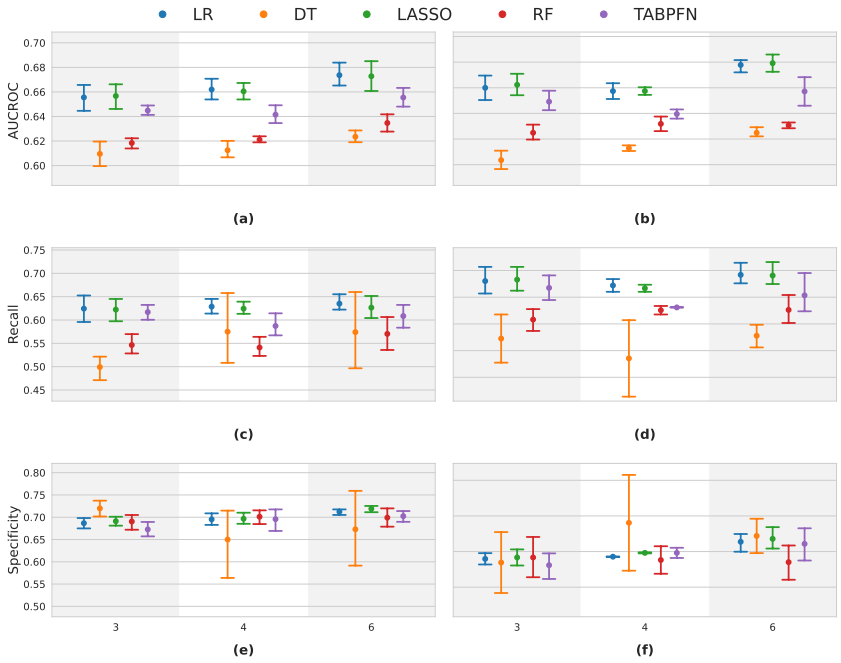


Figure 4. Nocturnal hypoglycemia (NH) prediction performance using random undersampling (RUS) and configurations with 3, 4, and 6 windows. Left panels show non-overlapping windows, and right panels show overlapping windows. Panels (**a**–**f**) display the area under the receiver operating characteristic curve (AUCROC), recall, and specificity, respectively. Points represent the mean performance for each classifier and window configuration, and bars indicate confidence intervals around the mean. Colors identify the classifiers, while alternating shaded vertical bands facilitate visual distinction between the different numbers of windows.

Figure 4c presents recall values for the non-overlapping windows. LASSO, LR, and TabPFN exhibited stable recall values across window configurations, whereas RF showed a small increase in recall at six- compared to three-window configurations. DT showed larger standard-deviation values, particularly at the four-window configuration. For overlapping windows (see Figure 4d), recall values tend to be higher, with the six-window configuration yielding the highest mean recall for most classifiers. TabPFN consistently exhibited low variability across random seeds in both overlapping and non-overlapping windows.

Figure 4e,f presents the specificity results for the non-overlapping and overlapping windows. For non-overlapping windows (see Figure 4e), LR and LASSO consistently achieved high specificity across all window configurations, with a slight improvement observed for the six-window configuration. RF and TabPFN exhibited similar trends, whereas DT showed greater variability across random seeds, particularly for configurations with more windows. For overlapping windows (see Figure 4f), specificity followed a similar pattern, with the six-window configuration yielding higher mean values for most classifiers.

5.3. Ensemble Feature Selection for Addressing High Dimensionality

To address the high dimensionality, ensemble FS was employed using Relief as the base selector, with feature rankings aggregated through majority voting. To isolate the effect of FS from class imbalance, all experiments were conducted with a fixed IR of 0.6. By keeping the IR constant, the influence of class distribution on feature relevance is controlled, allowing a focused assessment of the stability and predictive relevance of the selected features. The experiments were conducted using a non-overlapping six-window configuration and four ML models: DT, LASSO, RF, and LR. For each model, the FS process was repeated using

five random seeds, with the proportion of selected features ranging from 10% to 100% in 10% increments. TabPFN was not included in this analysis to limit the computational cost associated with evaluating multiple feature-retention percentages and random seeds.

Figure 5 shows the NH prediction results (measured by AUCROC) for different percentages of selected features. LR and LASSO achieved the highest AUCROC values across the evaluated feature subsets, followed by RF, whereas DT consistently exhibited the lowest performance. LR and LASSO also showed low variability across random seeds, while RF and DT presented greater variability for some feature subset sizes. For each feature subset, hyperparameters were optimized using grid search with three-fold cross-validation on the training set, followed by evaluation on the test set. Across the evaluated retention percentages, LR achieved its highest mean test-set AUCROC at 40% feature retention (0.6718), compared with 0.6667 when all features were used. This configuration reduced the feature set from 236 to 94 features and corresponded to the non-overlapping setting. Accordingly, subsequent feature-selection experiments used the 40% retention level. However, because this proportion was selected based on test-set performance, the resulting comparisons should be considered exploratory, as the selected retention level was not evaluated on an independent dataset.

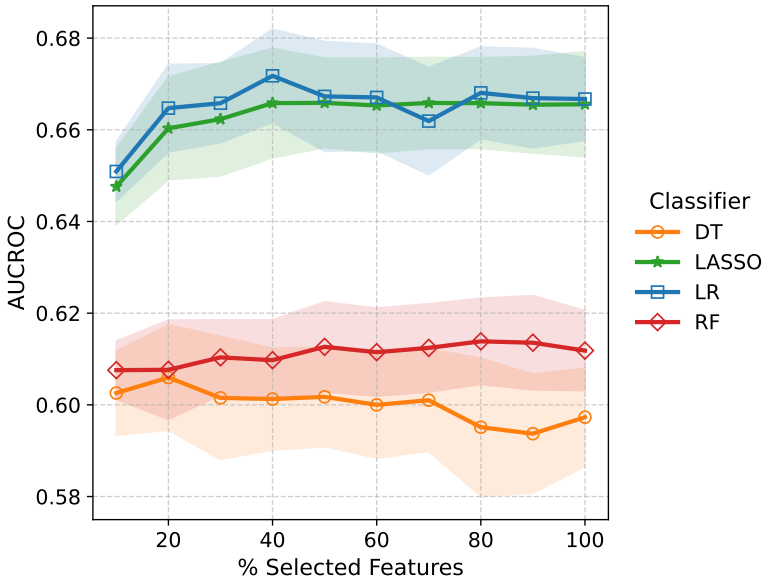


Figure 5. Nocturnal hypoglycemia (NH) prediction performance as a function of the percentage of features retained by bootstrap-based Relief voting. Experiments use random undersampling (RUS), a target imbalance ratio (IR) of 0.6, and a non-overlapping 6-window configuration. Feature subsets range from 10% to 100% in 10% increments, with five random seeds evaluated for each classifier. Colors identify the classifiers, and predictive performance is measured by AUCROC.

5.4. Comparison of Nocturnal Hypoglycemia Prediction Models with and Without Ensemble Feature Selection

This subsection presents the results of NH prediction with and without ensemble FS across different IRs. The analysis is presented separately for RUS and SMOTE. LASSO was excluded from this comparison due to convergence difficulties encountered during training with SMOTE-generated synthetic samples, which prevented its consistent evaluation across the different IRs.

For SMOTE, Figures 6 and 7 present the results for overlapping and non-overlapping windows, respectively. Note that the left panels show the evolution of the AUCROC as a function of the IR, whereas the right panels summarize the corresponding AUCROC distributions across all evaluated IR values using boxplots. Figure 6 presents the performance of

four classifiers for NH prediction using SMOTE resampling with overlapping windows, evaluated with (panels c,d) and without (panels a,b) ensemble FS. Overall, LR consistently achieved the highest predictive performance, maintaining an AUCROC of approximately 0.66 without FS and improving to 0.68 after FS, while also exhibiting the lowest performance variability across IRs. RF ranked second, with AUCROC increasing gradually from approximately 0.61 to 0.63 as the IR increased, and showing a more pronounced improvement when combined with FS. TabPFN achieved stable but comparatively lower performance, with AUCROC remaining close to 0.60–0.61 across all IRs. DT yielded the lowest predictive performance and showed the greatest variability. These results suggest that the effects of ensemble FS are classifier-dependent, with performance improvements observed for LR and RF, while showing a limited impact on DT and TabPFN.

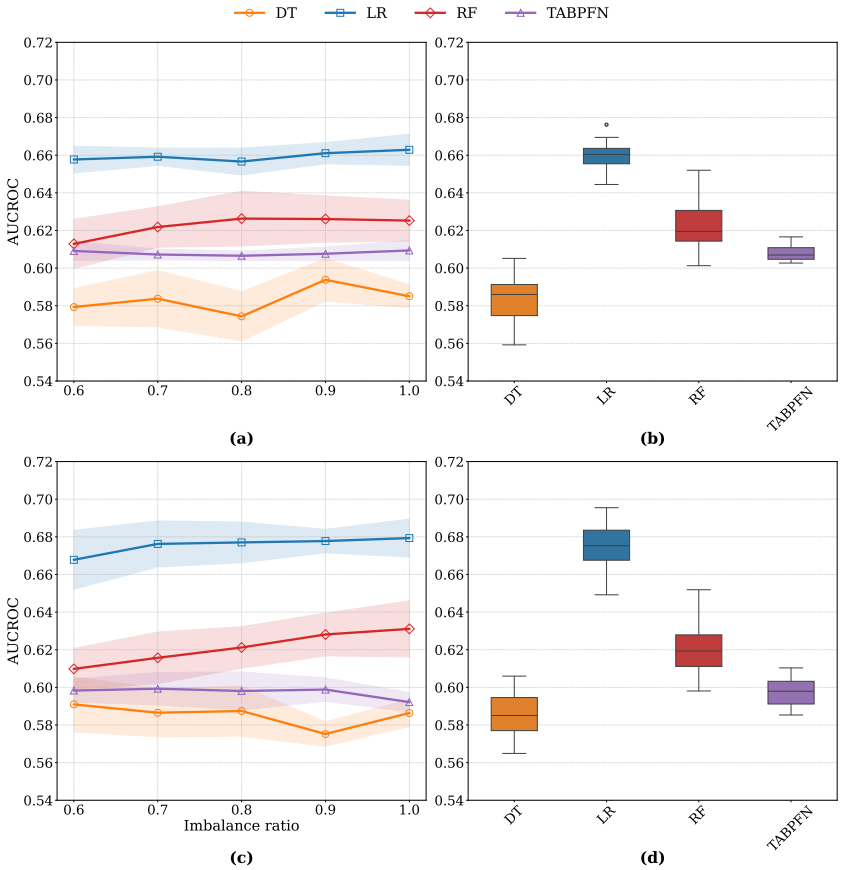


Figure 6. AUCROC for nocturnal hypoglycemia (NH) prediction using the Synthetic Minority Over-sampling Technique (SMOTE) and overlapping windows. Panels (**a**,**b**) show models without ensemble feature selection (FS), whereas panels (**c**,**d**) show models retaining the top 40% of features ranked by bootstrap-based Relief voting. Panels (**a**,**c**) display performance across target training imbalance ratios (minority-to-majority sample ratios) of 0.6–1.0. Panels (**b**,**d**) summarize performance across these ratios using boxplots. Colors identify the classifiers.

Figure 7 presents the performance of the four classifiers for NH prediction using SMOTE resampling with non-overlapping windows, evaluated without (panels a,b) and with (panels c,d) ensemble FS. Similar to the overlapping-window approach, LR achieved the highest AUCROC across the evaluated IRs and improved after ensemble FS. RF ranked second and improved as the IR increased, both with and without FS, whereas DT and TabPFN showed comparable performance. After applying FS, DT exhibited a slight performance improvement, surpassing TabPFN at the highest IR. Overall, the non-overlapping-window strategy preserved the superior performance of LR observed with

overlapping windows, whereas RF benefited more noticeably from progressively balanced class distributions.

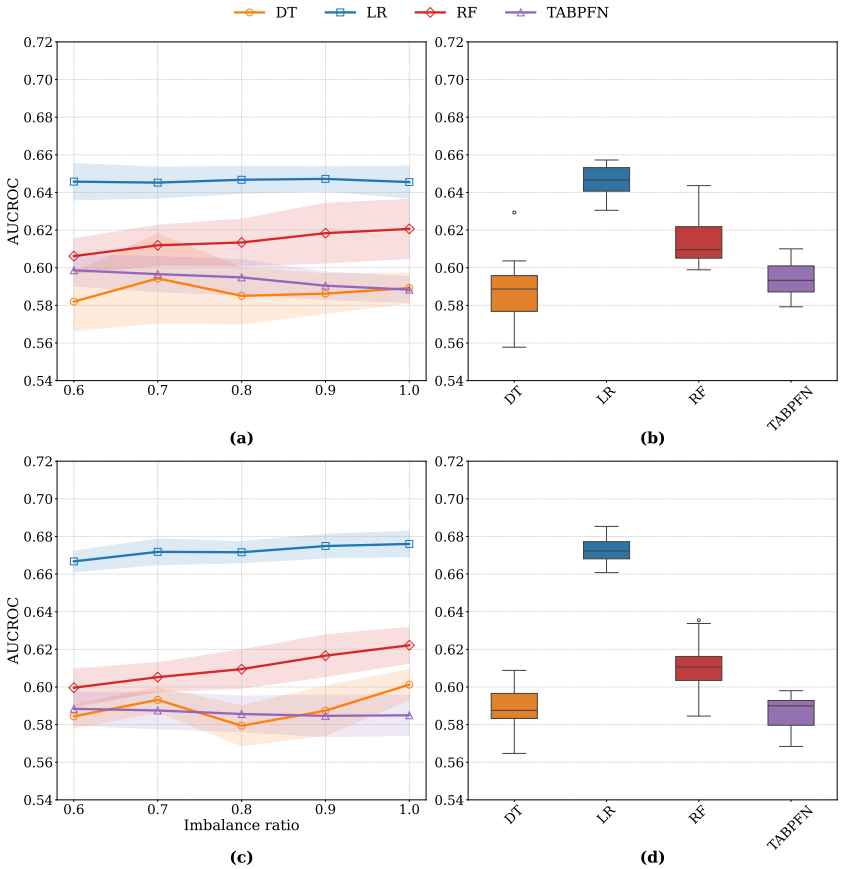


Figure 7. AUCROC for nocturnal hypoglycemia (NH) prediction using the Synthetic Minority Oversampling Technique (SMOTE) and non-overlapping windows. Panels (**a**,**b**) show models without ensemble feature selection, whereas panels (**c**,**d**) show models retaining the top 40% of features ranked by bootstrap-based Relief voting. Panels (**a**,**c**) display performance across target training imbalance ratios (minority-to-majority sample ratios) of 0.6–1.0. Panels (**b**,**d**) summarize performance across these ratios using boxplots. Colors identify the classifiers.

For RUS, Figures 8 and 9 present the results for overlapping and non-overlapping windows, respectively. Figure 8 presents the performance of the four classifiers for NH prediction using RUS with overlapping windows, evaluated without (panels a,b) and with (panels c,d) ensemble FS. LR achieved the highest predictive performance, with AUCROC increasing from approximately 0.67 to 0.68 as the imbalance ratio (IR) increased. TabPFN emerged as the second-best performing model, representing a substantial improvement over its performance with SMOTE. Its AUCROC increased steadily from approximately 0.64 to 0.67, and the corresponding boxplots showed a higher median performance than both RF and DT. RF and DT achieved similar results, with AUCROC increasing gradually as the IR increased. Applying ensemble FS resulted in minor changes to the overall ranking and performance of the classifiers, with LR and TabPFN maintaining a clear advantage over RF and DT.

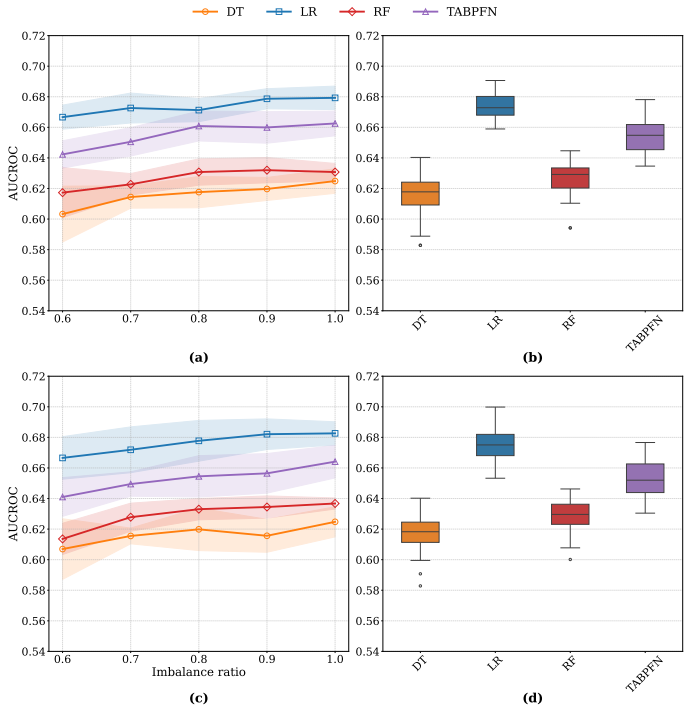


Figure 8. AUCROC for nocturnal hypoglycemia (NH) prediction using random undersampling (RUS) and overlapping windows. Panels (**a**,**b**) show models without ensemble feature selection, whereas panels (**c**,**d**) show models retaining the top 40% of features ranked by bootstrap-based Relief voting. Panels (**a**,**c**) display performance across target training imbalance ratios (minority-to-majority sample ratios) of 0.6–1.0. Panels (**b**,**d**) summarize performance across these ratios using boxplots. Colors identify the classifiers.

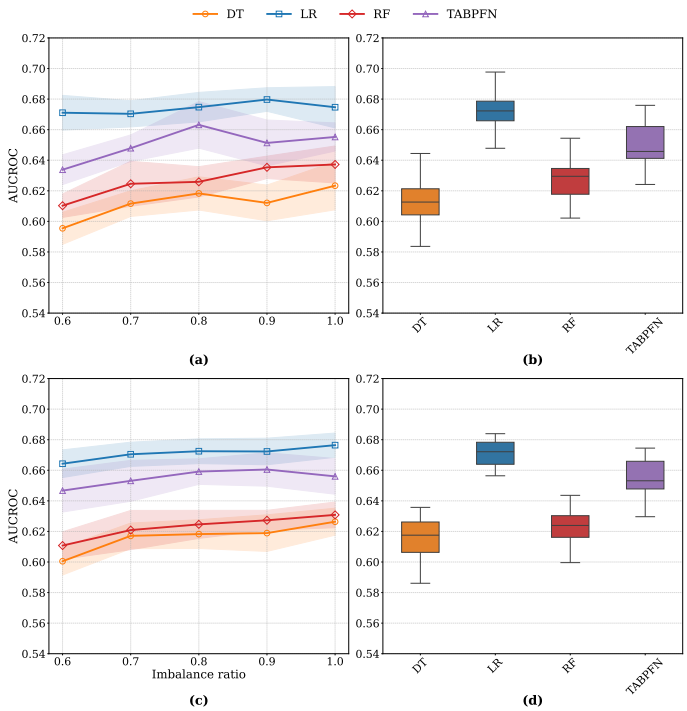


Figure 9. AUCROC for nocturnal hypoglycemia (NH) prediction using random undersampling (RUS) and non-overlapping windows. Panels (**a**,**b**) show models without ensemble feature selection, whereas panels (**c**,**d**) show models retaining the top 40% of features ranked by bootstrap-based Relief voting. Panels (**a**,**c**) display performance across target training imbalance ratios (minority-to-majority sample ratios) of 0.6–1.0. Panels (**b**,**d**) summarize performance across these ratios using boxplots. Colors identify the classifiers.

Figure 9 presents the performance of the four classifiers for NH prediction using RUS with non-overlapping windows, evaluated without (panels a,b) and with (panels c,d) ensemble FS. Similar to the results obtained with overlapping windows, LR achieved the highest AUCROC across the evaluated IRs. TabPFN again ranked second, maintaining the trend observed with overlapping windows. TabPFN showed an initial increase in AUCROC followed by a slight decline at the highest IRs, while RF and DT showed comparable performance and an increasing trend. Similar to the overlapping-window results, applying ensemble FS had only a limited effect on classifier performance, preserving the overall ranking of the models.

5.5. SHAP Analysis Using CGM-Derived Features

A SHAP-based interpretability analysis was performed to identify the contribution of CGM-derived features for NH prediction. Two scenarios were analyzed: models trained with (i) all features and (ii) the top-k features identified by ensemble FS.

Figure 10 shows the SHAP-based feature ranking for NH prediction. Panels (a,b) show the LR model trained with RUS and overlapping windows using all features. As shown, the most informative features are predominantly related to glucose exposure and central tendency measures computed over overlapping nested windows. In the SHAP beeswarm representation (see Figure 10a), `auc|5`, `mean|4`, `auc|3`, and `auc|4` were identified as the most relevant features for NH prediction. Higher values of `auc|5`, `auc|3`, and `auc|4` were predominantly associated with positive SHAP contributions, whereas higher values of `mean|4` were predominantly associated with negative contributions. Thus, these features contributed in different directions to the model output rather than indicating a uniform association between higher glucose values and higher predicted NH risk.

In contrast, features extracted from earlier time windows, including `auc|0`, `mean|1`, and `auc|1`, exhibited SHAP values that were closely centered around zero in Figure 10a, indicating a limited contribution to model predictions. Several other displayed features had smaller mean absolute SHAP contributions than the leading mean-glucose and AUC features.

In the second scenario, we considered LR with RUS and ensemble FS (integrating Relief and voting), retaining the top 40% of features. Figure 10c,d presents the corresponding SHAP values for this scenario. The same group of features, particularly `mean|4`, `auc|4`, `auc|5`, and `auc|3`, remained the most relevant for NH prediction. However, their contributions exhibited reduced redundancy and clearer separation in SHAP value distributions. The leading features maintained mean absolute SHAP values above 1.0, while features with limited predictive utility exhibited substantially reduced contributions. Furthermore, several features demonstrated reduced SHAP variability across individual samples, suggesting a more stable and consistent contribution pattern. This suggests that ensemble FS may help reduce feature dimensionality while retaining predictive information within a smaller subset of CGM-derived features.

In both scenarios, several features from windows three to five had some of the largest mean absolute SHAP contributions. These findings identify influential metric–window combinations within the evaluated models, without establishing a physiological explanation for their associations with NH.

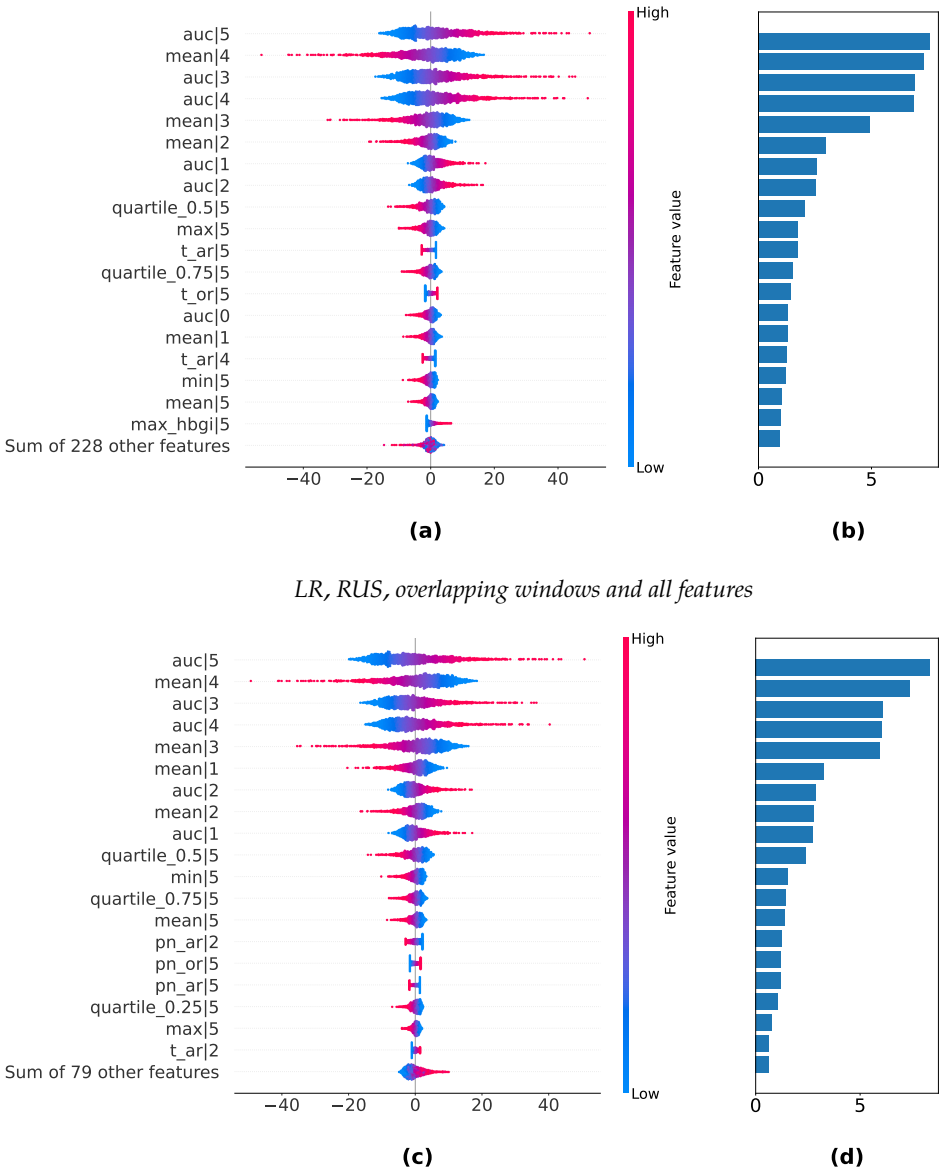


Figure 10. Shapley additive explanations (SHAP)-based interpretation of logistic regression (LR) models trained using random undersampling (RUS) and overlapping windows. Panels (**a**,**b**) show the model using all features, whereas panels (**c**,**d**) show the model retaining the top 40% selected by bootstrap-based Relief voting. Beeswarm plots (**a**,**c**) display sample-level feature contributions, with color indicating feature value from low (blue) to high (red). Bar plots (**b**,**d**) summarize mean absolute SHAP values across the analyzed samples. Feature-name suffixes indicate temporal window indices.

6. Discussion

This study proposed a data-driven approach combining ensemble FS synthetic data augmentation techniques to address high dimensionality and class imbalance in NH prediction. The predictive performance achieved in this study (AUCROC $\approx$ 0.68) represents a moderate yet clinically meaningful level of accuracy for NH prediction. Our work achieved predictive performance comparable to recent time-series classification models in the literature [26]. These results characterize performance across the temporal representations, FS settings, and resampling strategies evaluated within our cohort.

The performance of LR may suggest that the predictive information contained in the CGM summary features is largely captured by additive feature combinations. The

hyperparameter search included both L1 and L2 regularization [42], which is consistent with the competitive performance observed for LR and LASSO. However, the present results do not establish either linearity of the underlying relationships or the specific contribution of regularization to model performance. RF and TabPFN did not consistently outperform LR, while TabPFN performed better with RUS than with SMOTE, suggesting sensitivity to the resampling strategy. Overall, these findings neither demonstrate overfitting nor support attributing differences in performance only to model complexity.

The classifier ranking also differs from that reported by Piersanti et al. [15], whose supplementary analysis showed higher AUC values for RF than for LR across both temporal configurations. Therefore, the relative advantage of LR observed in our study does not appear to generalize across NH prediction studies. Differences in study population, feature representations, and evaluation protocols prevent attributing this discrepancy to any single factor.

Piersanti et al. [15] represents the closest methodological reference to our work. However, our study extends the prediction setting beyond exercise-associated nights by including adults experiencing NH irrespective of preceding exercise and by evaluating both overlapping and non-overlapping temporal window strategies. A direct comparison of bootstrap-based Relief voting and VIF–mRMR under a common evaluation protocol would be required to determine whether either FS strategy offers a predictive advantage. In comparison with Roquemen-Echeverri et al. [28], our models are based exclusively on CGM-derived information. Therefore, differences in feature relevance should be interpreted in light of both the modeling strategy and the information available to each model. Our study also addresses a distinct objective from Herrero et al. [29]: rather than estimating the potential glycemic effects of acting on predicted risk, we focus on the discrimination of future NH events.

A notable finding of this study is the influence of temporal representation on predictive performance. Extending the number of windows resulted in more robust and consistent classifier performance, with the six-window configuration achieving the most balanced combination of AUCROC, recall, and specificity. Moreover, overlapping windows consistently improved recall without adversely affecting specificity, suggesting that overlapping windows may provide complementary representations of glucose dynamics for NH prediction. Piersanti et al. [15] reported higher AUCROC with SMOTE in both temporal configurations, although sensitivity and balanced accuracy varied. In our experiments, LR achieved approximately 0.68 with SMOTE or RUS, whereas TabPFN favored RUS (Figures 6–9). Thus, SMOTE yielded LR performance comparable to that obtained with RUS, but its relative performance varied across classifiers. Because no condition without resampling was evaluated in these experiments, the present comparisons do not establish whether SMOTE improves performance over training without resampling. From a clinical perspective, achieving high sensitivity is essential to ensure the reliable detection of true hypoglycemic events and facilitate timely intervention.

By integrating model interpretability techniques with window-based CGM feature extraction, we identified the most informative features for NH prediction. Window-based extraction yielded over 200 candidate features. Bootstrap-based Relief voting combined selections across resampled training subsets into a consensus ranking to reduce single-sample dependence; redundancy reduction and SHAP-explanation stability improvements remain unestablished. SHAP revealed that CGM-derived features related to aggregated glucose exposure, such as mean glucose and AUC-based metrics from later windows (e.g., $auc|4$, $auc|5$), are relevant for NH prediction.

These temporally localized summaries are consistent with Piersanti et al.'s [15] finding that predictive glycemic information spans both exercise and recovery periods. However, this finding supports the relevance of temporal representation rather than a shared physiological mechanism. Our analysis windows were not aligned with exercise events and therefore cannot establish post-exercise causality. However, the overlapping windows vary simultaneously in two dimensions: their proximity to the prediction point and their duration. Consequently, SHAP rankings cannot determine whether the observed importance is driven by temporal proximity or by window length. Therefore, these rankings should not be interpreted as evidence that one time interval is necessarily more clinically relevant than another.

Limitations and Future Work

This study has several limitations. First, the cohort included 52 individuals with T1D recruited from a single hospital, and no external validation was performed. Training and test sets were separated at the participant–day level, allowing different days from the same participant to appear in both subsets. Dependence between observations from the same individual may therefore contribute to the reported performance, which does not establish generalization to previously unseen participants. Furthermore, the split was not designed as a chronological evaluation of future observations. Future studies should assess performance using participant-independent and temporally appropriate validation designs, as well as larger, independent multicenter cohorts to evaluate transportability [43].

Second, the models relied exclusively on CGM-derived features. Insulin dosing, carbohydrate intake, and physical activity were not included as features. Incorporating these variables may improve discrimination beyond the moderate AUCROC obtained in this study and it would make model explanations more clinically actionable. Additionally, CGM-derived metrics were computed exclusively using GlucoStats, and no cross-package comparison was performed to assess the consistency of feature values across different software implementations.

Third, because the nested windows differ simultaneously in duration and temporal coverage while sharing a common endpoint, feature importance cannot be interpreted exclusively as an effect of temporal proximity to the prediction time. Future work could use windows of equal length to disentangle window duration from temporal proximity and support a clearer temporal interpretation of feature importance.

A potential clinical application would be to present an NH prediction together with the underlying CGM trace and the features contributing to that prediction. This approach could direct clinical review toward the glucose patterns associated with the predicted risk. Interpretation would also require contextual information, including insulin administration, meals, and physical activity, as emphasized in the GlucoStats paper [31]. These explanations could support the clinical assessment of a prediction. However, our analysis does not establish which treatment changes would reduce NH risk. Prediction-guided preventive interventions have been explored in previous work, including a computer simulation study that evaluated bedtime carbohydrate recommendations [20], but their effectiveness was not assessed in the present study.

7. Conclusions

This study evaluated a framework integrating ensemble FS, resampling, and window-based feature extraction for NH prediction. The proposed methodology addresses two key challenges in AI-based NH prediction: class imbalance and the high dimensionality of CGM-derived features. Building on prior evidence from post-exercise NH prediction [15], this study extends the evaluation of temporal CGM representations to an adult setting not

restricted to exercise-associated nights. Within this cohort, the six-window configuration yielded the highest reported AUCROC. For the two individual LR models examined through SHAP, test-set AUCROC values calculated from continuous prediction scores were approximately 0.74 with both the full and selected feature sets. These results do not establish an advantage of FS, as the models also used different resampling ratios. SHAP analysis identified the CGM-derived features and temporal representations with the largest contributions to the evaluated models' NH predictions. This work provides an interpretable framework for the early prediction of NH, supporting the development of clinically applicable decision-support tools for individuals with T1D. By enabling timely identification of elevated NH risk, the proposed framework may facilitate preventive therapeutic adjustments and contribute to improving patient safety and confidence in diabetes management.

Author Contributions: Conceptualization, D.C.-M., C.S.-R. and B.P.-A.; methodology, D.C.-M.; software, D.C.-M., F.J.L.-A. and B.P.-A.; validation, C.B.A., A.M.W. and C.G.; formal analysis, B.P.-A.; investigation, B.P.-A.; data curation, C.B.A., F.J.L.-A. and B.P.-A.; visualization, B.P.-A. and C.S.-R.; writing—original draft preparation, B.P.-A.; writing—review and editing, D.C.-M., F.J.L.-A., C.B.A., A.M.W., C.G. and C.S.-R.; supervision, D.C.-M. and C.S.-R.; project administration, C.S.-R. and C.G.; funding acquisition, C.S.-R. and C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Commission through the H2020-EU.3.1.4.2, European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under Grant Agreement 101017385; and by the Spanish Government by the Grant AAVis-BMR PID2019-107768RA-I00/AEI/10.13039/50110 0011033 and through HEALTHtray under the grant PID2023-149457OB-I00. The study sponsors have not been involved in any stage of the study.

Institutional Review Board Statement: The Provincial Ethics Committee of Las Palmas determined that the study was exempt from ethical review, as it did not fall under the scope of Biomedical Research regulations.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used in this study contain confidential patient health information; however, it may be made available upon reasonable request.

Acknowledgments: The authors thank Pablo Peiro-Corbacho for their insightful comments and technical feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUCROC	Area Under the Receiver Operating Characteristic Curve
CGM	Continuous Glucose Monitoring
CIP	Class Imbalance Problem
DT	Decision Tree
FS	Feature Selection
IR	Imbalance Ratio
LR	Logistic Regression
ML	Machine Learning
NH	Nocturnal Hypoglycemia
RF	Random Forest
ROC	Receiver Operating Characteristic
RUS	Random Under-Sampling
SHAP	SHapley Additive exPlanations

SMOTE	Synthetic Minority Over-sampling Technique
TabPFN	Tabular Prior-Data Fitted Network
TIR	Time In Range
VIF-mRMR	Variance Inflation Factor–Minimum Redundancy Maximum Relevance
XAI	Explainable Artificial Intelligence

References

1. Gregory, G.A.; Robinson, T.I.; Linklater, S.E.; Wang, F.; Colagiuri, S.; de Beaufort, C.; Donaghue, K.C.; Harding, J.L.; Wander, P.L.; Zhang, X.; et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: A modelling study. *Lancet Diabetes Endocrinol.* **2022**, *10*, 741–760. Correction in *Lancet Diabetes Endocrinol.* **2022**, *10*, e11. [CrossRef] [PubMed]
2. Yale, J. Nocturnal hypoglycemia in patients with insulin-treated diabetes. *Diabetes Res. Clin. Pract.* **2004**, *65*, S41–S46. [CrossRef]
3. Zhu, B.; Abu-Irsheid, G.M.; Martyn-Nemeth, P.; Reutrakul, S. Type 1 Diabetes, Sleep, and Hypoglycemia. *Curr. Diabetes Rep.* **2021**, *21*, 55. [CrossRef]
4. Klonoff, D.C.; Ahn, D.; Drincic, A. Continuous glucose monitoring: A review of the technology and clinical use. *Diabetes Res. Clin. Pract.* **2017**, *133*, 178–192. [CrossRef]
5. Lara-Abelenda, F.J.; Chushig-Muzo, D.; Peiro-Corbacho, P.; Wägner, A.M.; Granja, C.; Soguero-Ruiz, C. Personalized glucose forecasting for people with type 1 diabetes using large language models. *Comput. Methods Programs Biomed.* **2025**, *265*, 108737. [CrossRef]
6. Lara-Abelenda, F.J.; Chushig-Muzo, D.; Wägner, A.M.; Tayefi, M.; Soguero-Ruiz, C. Interpretable and Multimodal Fusion Methodology to Predict Severe Hypoglycemia in Adults with Type 1 Diabetes. *Eng. Appl. Artif. Intell.* **2025**, *144*, 110142. [CrossRef]
7. Chushig-Muzo, D.; Calero-Díaz, H.; Fabelo, H.; Årsand, E.; van Dijk, P.R.; Soguero-Ruiz, C. Characterizing the Impact of Physical Activity on Patients with Type 1 Diabetes Using Statistical and Machine Learning Models. *Appl. Sci.* **2024**, *14*, 9870. [CrossRef]
8. Darpit, D.; Vyas, K.; Jayagopal, J.K.; Garcia, A.; Erraguntla, M.; Lawley, M. A personalized federated learning-based glucose prediction algorithm for high-risk glycemic excursion regions in type 1 diabetes. *Sci. Rep.* **2025**, *15*, 38376. Correction in *Sci. Rep.* **2026**, *16*, 5742. [CrossRef]
9. Mosquera-Lopez, C.; Wilson, L.M.; El Youssef, J.; Hilts, W.; Leitschuh, J.; Branigan, D.; Gabo, V.; Eom, J.H.; Castle, J.R.; Jacobs, P.G. Enabling fully automated insulin delivery through meal detection and size estimation using artificial intelligence. *npj Digit. Med.* **2023**, *6*, 39. [CrossRef]
10. Battelino, T.; Danne, T.; Bergenstal, R.M.; Amiel, S.A.; Beck, R.; Biester, T.; Bosi, E.; Buckingham, B.A.; Cefalu, W.T.; Close, K.L.; et al. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range. *Diabetes Care* **2019**, *42*, 1593–1603. [CrossRef] [PubMed]
11. Rodbard, D. Clinical Interpretation of Indices of Quality of Glycemic Control and Glycemic Variability. *Postgrad. Med.* **2011**, *123*, 107–118. [CrossRef]
12. Duckworth, C.; Guy, M.J.; Kumaran, A.; O'Kane, A.A.; Ayobi, A.; Chapman, A.; Marshall, P.; Boniface, M. Explainable Machine Learning for Real-Time Hypoglycemia and Hyperglycemia Prediction and Personalized Control Recommendations. *J. Diabetes Sci. Technol.* **2024**, *18*, 113–123. [CrossRef] [PubMed]
13. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
14. Liu, Y.; Guo, C.; Li, X.; Li, S.; Huang, J.; Zhao, L.; Zhu, Y.; Liu, X.; Wang, B.; Lin, R.; et al. Machine learning-based online prediction of nocturnal hypoglycemia in elderly patients with type 2 diabetes. *Front. Endocrinol.* **2026**, *16*, 1685969. [CrossRef]
15. Piersanti, A.; Manes, G.M.; Del Giudice, L.L.; Burattini, L.; Göbl, C.; Tura, A.; Morettini, M. Prediction of Nocturnal Hypoglycemia Following Exercise in Type 1 Diabetes Using Temporally Structured CGM-Derived Digital Biomarkers. *Sensors* **2026**, *26*, 3842. [CrossRef]
16. Vu, L.H.; Kefayati, S.; Idé, T.; Pavuluri, V.N.; Jackson, G.P.; Latts, L.; Zhong, Y.; Agrawal, P.; Chang, Y.C. Predicting Nocturnal Hypoglycemia from Continuous Glucose Monitoring Data with Extended Prediction Horizon. In *Proceedings of the AMIA Annual Symposium Proceedings*; American Medical Informatics Association: Washington, DC, USA, 2019; pp. 874–882.
17. 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]

18. Jensen, M.H.; Dethlefsen, C.; Vestergaard, P.; Hejlesen, O. Prediction of Nocturnal Hypoglycemia From Continuous Glucose Monitoring Data in People with Type 1 Diabetes: A Proof-of-Concept Study. *J. Diabetes Sci. Technol.* **2020**, *14*, 250–256. [CrossRef]
19. Berikov, V.B.; Kutnenko, O.A.; Semenova, J.F.; Klimontov, V.V. Machine Learning Models for Nocturnal Hypoglycemia Prediction in Hospitalized Patients with Type 1 Diabetes. *J. Pers. Med.* **2022**, *12*, 1262. [CrossRef]
20. Mosquera-Lopez, C.; Roquemen-Echeverri, V.; Tyler, N.S.; Patton, S.R.; Clements, M.A.; Martin, C.K.; Riddell, M.C.; Gal, R.L.; Gillingham, M.; Wilson, L.M.; et al. Combining Uncertainty-Aware Predictive Modeling and a Bedtime Smart Snack Intervention to Prevent Nocturnal Hypoglycemia in People with Type 1 Diabetes on Multiple Daily Injections. *J. Am. Med. Inform. Assoc.* **2023**, *31*, 109–118. [CrossRef]
21. Kozinetz, R.M.; Berikov, V.B.; Semenova, J.F.; Klimontov, V.V. Machine Learning and Deep Learning Models for Nocturnal High- and Low-Glucose Prediction in Adults with Type 1 Diabetes. *Diagnostics* **2024**, *14*, 740. [CrossRef]
22. Yu, X.; Yang, Z.; Wang, X.; Sun, X.; Shen, R.; Li, H.; Zhang, M. A Prior-Knowledge-Guided Dynamic Attention Mechanism to Predict Nocturnal Hypoglycemic Events in Type 1 Diabetes. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 378. [CrossRef]
23. Leutheuser, H.; Bartholet, M.; Marx, A.; Pfister, M.; Burckhardt, M.A.; Bachmann, S.; Vogt, J.E. Predicting Risk for Nocturnal Hypoglycemia after Physical Activity in Children with Type 1 Diabetes. *Front. Med.* **2024**, *11*, 1439218. [CrossRef]
24. Voegeli, M.; Laguna, S.; Leutheuser, H.; Pfister, M.; Burckhardt, M.A.; Vogt, J.E. Beyond Glucose-Only Assessment: Advancing Nocturnal Hypoglycemia Prediction in Children with Type 1 Diabetes. *arXiv* **2025**, arXiv:2504.09299.
25. Afentakis, I.; Unsworth, R.; Herrero, P.; Oliver, N.; Reddy, M.; Georgiou, P. Development and Validation of Binary Classifiers to Predict Nocturnal Hypoglycemia in Adults with Type 1 Diabetes. *J. Diabetes Sci. Technol.* **2025**, *19*, 153–160. [CrossRef] [PubMed]
26. Lara-Abelenda, F.J.; Chushig-Muzo, D.; Acosta, C.B.; Wägner, A.M.; Granja, C.; Soguero-Ruiz, C. Evaluating Time Series Classification Models for Nocturnal Hypoglycemia: From Predictive Performance to Environmental Impact. *IEEE Access* **2025**, *13*, 150756–150771. [CrossRef]
27. Tziavaras, K.; Athanasiou, M.; Nikita, K.S. Development of autoencoder-guided attention-LSTM models for predicting nocturnal hypoglycemia risk in Type 1 Diabetes. *IEEE Trans. Biomed. Eng.* **2026**, 1–14. [CrossRef]
28. Roquemen-Echeverri, V.; Jacobs, P.G.; Wilson, L.M.; Pinsonault, J.; Branigan, D.; Eom, J.; Chen, D.; Ling, H.; Aby-Daniel, D.; Chen, K.; et al. Explainable AI-based Decision Support for Nocturnal Hypoglycemia Prevention in Type 1 Diabetes. *arXiv* **2026**, arXiv:2608.21563. [CrossRef]
29. Herrero, P.; Andorrà, M.; Breton, M.D.; Peak, A.; Koehler, M.; Klopfenstein, Y.; Leppäaho, E.; Zanon, M.; Ringemann, C.; Lustenberger, P.; et al. Glucose Predictions Improve Glycemic Control: A Digital Twin Evaluation. *Diabetes Technol. Ther.* **2026**, *28*, 1052–1062. [CrossRef] [PubMed]
30. WARIFA Project. WARIFA: Personalized Risk Prediction for Non-Communicable Diseases. 2026. Available online: https://www.warifa.eu/ (accessed on 30 July 2026).
31. Peiro-Corbacho, P.; Lara-Abelenda, F.J.; Chushig-Muzo, D.; Wägner, A.M.; Granja, C.; Soguero-Ruiz, C. Glucostats: An Efficient Python Library for Glucose Time Series Feature Extraction and Visual Analysis. *BMC Bioinform.* **2025**, *26*, 230. [CrossRef]
32. Broll, S.; Urbanek, J.; Buchanan, D.; Chun, E.; Muschelli, J.; Punjabi, N.M.; Gaynanova, I. Interpreting blood GLUcose data with R package iglu. *PLoS ONE* **2021**, *16*, e0248560. [CrossRef]
33. Vigers, T.; Chan, C.L.; Snell-Bergeon, J.; Bjornstad, P.; Zeitler, P.S.; Forlenza, G.; Pyle, L. cgmanalysis: An R package for descriptive analysis of continuous glucose monitor data. *PLoS ONE* **2019**, *14*, e0216851. [CrossRef]
34. Piersanti, A.; Giurato, F.; Göbl, C.; Burattini, L.; Tura, A.; Morettini, M. Software Packages and Tools for the Analysis of Continuous Glucose Monitoring Data. *Diabetes Technol. Ther.* **2023**, *25*, 69–85. [CrossRef]
35. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
36. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [CrossRef]
37. Apicella, A.; Isgrò, F.; Prevete, R. Don't push the button! Exploring data leakage risks in machine learning and transfer learning. *Artif. Intell. Rev.* **2025**, *58*, 339. [CrossRef]
38. Nogueira, S.; Sechidis, K.; Brown, G. On the Stability of Feature Selection Algorithms. *J. Mach. Learn. Res.* **2018**, *18*, 1–54. [CrossRef]
39. Saeys, Y.; Abeel, T.; Van de Peer, Y. Robust Feature Selection Using Ensemble Feature Selection Techniques. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008*; Lecture Notes in Computer Science; Daelemans, W., Goethals, B., Morik, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5212, pp. 313–325.
40. Urbanowicz, R.J.; Meeker, M.; La Cava, W.; Olson, R.S.; Moore, J.H. Relief-based feature selection: Introduction and review. *J. Biomed. Inform.* **2018**, *85*, 189–203. [CrossRef]
41. Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput. Methods Programs Biomed.* **2022**, *214*, 106584. [CrossRef]

42. Friedman, J.H.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
43. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. Correction in *BMJ* **2024**, *385*, q902. [CrossRef] [PubMed]

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.