View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Northumbria Research Link

Northumbria Research Link

Citation: Rodríguez-Rodríguez, Ignacio, Rodríguez, José-Víctor, Woo, Wai Lok, Wei, Bo and Pardo-Quiles, Domingo-Javier (2021) A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. Applied Sciences, 11 (4). p. 1742. ISSN 2076-3417

Published by: MDPI

URL: https://doi.org/10.3390/app11041742 <https://doi.org/10.3390/app11041742>

This version was downloaded from Northumbria Research Link: http://nrl.northumbria.ac.uk/id/eprint/45441/

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: http://nrl.northumbria.ac.uk/policies.html

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

University Library

Northumbria
University
NEWCASTLE

applied sciences

MDPI

Article

A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus

Ignacio Rodríguez-Rodríguez [1,*], José-Víctor Rodríguez [2], Wai Lok Woo [3], Bo Wei [3] and Domingo-Javier Pardo-Quiles [2]

[1] Departamento de Ingeniería de Comunicaciones, Universidad de Málaga, 29071 Málaga, Spain
[2] Departamento de Tecnologías de la Información y las Comunicaciones, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain; jvictor.rodriguez@upct.es (J.-V.R.); domingo.pardo@upct.es (D.-J.P.-Q.)
[3] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK; wailok.woo@northumbria.ac.uk (W.L.W.); bo.wei@northumbria.ac.uk (B.W.)
* Correspondence: ignacio.rodriguez@ic.uma.es

Abstract: Type 1 diabetes mellitus (DM1) is a metabolic disease derived from falls in pancreatic insulin production resulting in chronic hyperglycemia. DM1 subjects usually have to undertake a number of assessments of blood glucose levels every day, employing capillary glucometers for the monitoring of blood glucose dynamics. In recent years, advances in technology have allowed for the creation of revolutionary biosensors and continuous glucose monitoring (CGM) techniques. This has enabled the monitoring of a subject's blood glucose level in real time. On the other hand, few attempts have been made to apply machine learning techniques to predicting glycaemia levels, but dealing with a database containing such a high level of variables is problematic. In this sense, to the best of the authors' knowledge, the issues of proper feature selection (FS)—the stage before applying predictive algorithms—have not been subject to in-depth discussion and comparison in past research when it comes to forecasting glycaemia. Therefore, in order to assess how a proper FS stage could improve the accuracy of the glycaemia forecasted, this work has developed six FS techniques alongside four predictive algorithms, applying them to a full dataset of biomedical features related to glycaemia. These were harvested through a wide-ranging passive monitoring process involving 25 patients with DM1 in practical real-life scenarios. From the obtained results, we affirm that Random Forest (RF) as both predictive algorithm and FS strategy offers the best average performance (Root Median Square Error, RMSE = 18.54 mg/dL) throughout the 12 considered predictive horizons (up to 60 min in steps of 5 min), showing Support Vector Machines (SVM) to have the best accuracy as a forecasting algorithm when considering, in turn, the average of the six FS techniques applied (RMSE = 20.58 mg/dL).

Keywords: diabetes mellitus; machine learning; feature selection; time series forecasting

Citation: Rodríguez-Rodríguez, I.; Rodríguez, J.-V.; Woo, W.L.; Wei, B.; Pardo-Quiles, D.-J. A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. *Appl. Sci.* **2021**, *11*, 1742. https://doi.org/10.3390/app11041742

Academic Editor: Pasi Fränti
Received: 8 January 2021
Accepted: 8 February 2021
Published: 16 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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 (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Type I diabetes mellitus (DM1) is generally accompanied by excessive blood sugar levels caused by the fact that the body is failing to create insulin. Within healthy subjects, blood glucose levels are regulated by glucose homeostasis, a closed-loop system [1]. The pancreas is the home of β cells that react to excessive glucose levels and create insulin to combat hyperglycemia. In DM1 subjects, such regulatory processes do not occur. DM1 is an autoimmune disease that causes the immune system to attack the pancreas' insu-

lin-producing cells. This is the most aggressive type of diabetes. DM1 subjects are incapable of producing insulin so they have to rely on either exogenous injections of the hormone or by wearing an insulin pump for regulation of glucose levels [2]. Management of diabetes aims to maintain homeostasis and to keep blood glucose at close to normal levels, alongside the avoidance of ketoacidosis, hypoglycemia, and additional longer term problems such as cardiovascular disease [3].

Normal levels of blood glucose are assumed to be between 80 and 120 mg/dL. Values below and above such a range would lead to hypoglycemia and hyperglycemia, respectively, and both situations could be inadvisable for the patient. Therefore, the desired interval is not too wide. In this sense, subjects with diabetes have to undertake frequent capillary blood glucose monitoring in order to maintain close to normoglycemia. Thankfully, the requirements for such monitoring have been removed from any patients with the advent of continuous glucose monitoring (CGM) technology, e.g., the Dexcom G6 (Dexcom Inc, San Diego, CA, USA). Such cutting-edge CGM technology is effective in reducing HbA1c (Hemoglobin A1C), achieving the HbA1c target and leading to reductions in fluctuations in glucose levels [4].

Despite the above, patients must also consider their diet and lifestyle to determine the amounts of insulin they require each day [5]. Basically, patients continuously make predictions as to where their blood glucose level will be in the future and the quantity and timing of insulin supplementation they will need to retain metabolic control in order to stave off hyperglycemia and hypoglycemia. Making an accurate self-forecast of blood glucose levels to set the correct dosage of insulin is a complex process. Patients must take a number of important elements into consideration and this can skew the subjective assessment of their needs result in errors. Thus, before being faced with an accurate calculation of insulin doses, it is necessary to achieve a glucose prediction that has an acceptable margin of error. The glucose prediction is therefore the first step.

The goal of a completely operative and functional artificial pancreas (AP) is an ambitious target to be reached. Thinking of a device capable of being autonomously maintain, in every situation, the blood glucose levels within a healthy range—free from hypo/hyperglycemia events—is nowadays far from being a tangible reality. Therefore, it is more reasonable to first consider intermediate steps, focusing on those aims which are more critical and, once they are under control, go further. In 2009, Kowalski suggested a pathway in the long trip to design an AP [6], in which one of the preliminary steps—before obtaining a proper calculation of the insulin doses that would be dispensed through an insulin pump—is undoubtedly to have a sufficiently accurate prediction. In fact, the absence of an accurate forecast of the evolution of glycaemia in a few minutes (taking into account past circumstances) could lead to errors in insulin dosage and, consequently, to poor management of blood glucose levels, which can be fatal.

It is fortunate that greater sophistication has been introduced into the technology for DM1 patients. Nowadays, advances in microcontrollers allow for the implementation of model predictive control (MPC) solutions that open the door to new options in the management of DM1, among other possibilities [7]. New wearable electronic technology may be a significant addition to the management scheme, providing a totally novel new prospect for managing diabetes and significantly improving the forecast accuracy of glucose levels. There are many forms of wearable technology that could be used for monitoring not only glucose levels but also a variety of physiological elements. Examples include taking electrocardiographs, monitoring heart rate, registering activity and exercise levels, and monitoring interstitial/blood glucose levels. There is a considerable amount of wearable technology that can register an individual's daily activities and physiological condition many times each minute. Using such resources, we have developed a full dataset of biomedical features related to glycaemia. We harvested these by undertaking an in-depth passive-monitoring program in the real world with 25 DM1 patients, helping us to effectively identify patterns which allow us to model the crucial elements of glucose levels.

With these data, we can then develop an accurate glycaemia forecast. Some very successful MPC-based solutions for glycaemia prediction, which make use of microcontrollers, have been presented in the literature [8]. Although time series values-prediction is well-trodden in the field of Machine Learning (ML), it is common to perform appropriate feature selection such as pre-processing, in order to enhance the efficacy of the predictive algorithm. Many of these predictive algorithms have only been applied to blood glucose prediction, but to the best of the authors' knowledge, there has been no comparative study not only of the prediction techniques, but also of the feature selection techniques. We find that it is complex to make fair comparisons with their results since each one uses different databases, with different added features, evaluating them in different ways (at different predictive horizons, different error evaluation metrics, etc.) and throughout different monitoring times. With all this, it is complicated to reach a conclusion on which one behaves better and it is not possible to extract overall conclusions. In addition, the previous phase of variable selection has not been compared either, so the influence of this on the improvement of the prediction has not been studied. The aim of the paper is therefore to make a fair comparison of these algorithms with a complete database. Moreover, the inclusions of the variables collected in our database are rarely seen in studies on blood glucose prediction.

Thus, in this paper, a rich and complete dataset has been acquired alongside a novel process of features selection that enhances the performance and comparison of the prediction has been developed. We believe that this paper will enable a confident combination of feature selection and forecasting techniques to achieve more precise predictions in an assumable predictive horizon (PH).

In short, the main contributions of this paper are as follows:

- A brief literature review on variable selection and prediction methods for glycaemia values in diabetics.
- To use an innovative database in the field of DM1, both in terms of the number of patients/variables considered and the monitoring time covered.
- To test different variable selection techniques.
- To combine these feature selection techniques with different predictive algorithms.
- To discuss the influence of the variable selection techniques on the performance of the predictive algorithm, as well as to study the accuracy achieved.

This research has employed a number of cutting-edge modeling and forecasting techniques. Implementation and analysis have been undertaken using six feature selection techniques alongside four forecasting techniques. The paper is organized as follows: Section 2 describes some previous works on forecasting glycaemia in DM1 to frame our research. In Section 3, we present the feature selection techniques and predictive algorithms. The monitoring campaign is depicted in Section 4, while Section 5 details the methodology followed in our work, with descriptions of the ML techniques. Section 6 offers the main results and discussion, and finally, we conclude the paper in Section 7.

2. Related Works

As mentioned in Section 1, it is crucial that an accurate range of variables and effective data collection methodology is employed when predicting blood glucose levels and that note is taken in the way they have previously been used. These variables must be related as time-series data, as the level of historical occurrences is considered as significant. Feature selection using a time-series is not the same as feature selection using standard data. With standard static data, target values only refer to the features' contemporary values. However, with feature selection in time-series, the target values are related to feature values at various points in the past as well as the contemporary value. This means that it is essential in feature selection in time-series to excise redundant/irrelevant variables and features, and selecting the relevant previous values for the creation of an effective dataset. In order to forecast glucose levels, it is widely recognized

that the effect of previous values has importance. One example is the research of Eskaf et al. [9] who found, by employing discrete Fourier transformations, that blood glucose levels vary as a result of meals within a single timeframe of eating. This means that there is the capacity (which is crucial) to select the correct influential variables and so reduce the dimensionality. This is an important stage in processing [10] prior to applying a data mining algorithm to any dataset.

Feature selection methodologies with time-series can be separated into filter, wrapper, and embedded techniques [11]. Filter techniques simply use the data in deciding on the features for retention. Wrapper techniques employ a learning algorithm wrapped around the feature search, selecting the features on the basis of its performance. With embedded methods, weighting is employed to control parameter values. In all cases, applications of variable selection methods can affect the course of diabetes mellitus. Balakrishnan et al. [12] applied Support Vector Machines (SVM) to rank variables affecting type 2 diabetes, and some hybrid methods have been proposed to optimize the diagnosis of diabetes [13]. However, there have not been many studies that apply variable selective methods to features that affect the immediate course of blood glucose in patients with type 1 DM. This could be due to some variables only recently being considered for predicting blood sugar [14]. In 2019, a sequential backward selection algorithm was successfully applied using a linear model in a cross-validation (CV) setting, obtaining an optimized and reduced subset [15]. Despite this, a study comparing the performance of different feature selection methods applied to diabetic patients' glycaemia has not been developed.

Numerous attempts have been made to develop a reliable prediction of glucose in DM1 patients. In this case, there have been approaches from a univariate point of view [16], using Autoregressive Integrated Moving Average (ARIMA), Random Forest (RF) and Support Vector Machines (SVM) with acceptable results. However, although univariate approximations can be interesting in computationally restricted environments, multivariate methods have demonstrated higher accuracy [17]. In this regard, some forecasting strategies need to be highlighted where results will be improved after a proper feature selection preliminary stage.

Linear Regression (LR) is probably the simplest methodology. This group of models endeavors to discover an assessment of the parameters of the model with the goal that the summation of the squared errors is minimized. Although it is the simplest, its accuracy could be sufficient, and due to its simplicity, it is easily executable even by limited hardware. In any case, recent approximations using Least Absolute Shrinkage and Selection Operator (LASSO) regression have achieved acceptable accuracy and good performance [18].

There are other methods using the same sort of techniques, for example, Gaussian Processes (GP) with Radial Basis Function Kernels (RBF) [19], which permit overall uniformity and limitless levels of basic functions. However, these are not often employed, although some researchers have used such techniques and the results have shown promise [20]. Some recent research has also looked at GP [21], examining the potential for automatic insulin delivery that could reduce the number of hypoglycemic events.

GP is a non-parametric methodology that revolves around creating a model of observable responses from a number of points in the training data (function values) and using them as multivariate normal random variables [22]. It is assumed that the function values will be distributed in such a way that the function will operate smoothly. Specifically, when closeness exists (in an Euclidean distance context) between matching input vectors that decay with divergences, the two function values will be closely correlated. Employing an assumed distribution by applying a basic probability manipulation allows for posterior distribution of hitherto unpredicted function values.

RF algorithms use a method referred to as bagging, which re-samples data instances a number of times in order to create several training subsets on the same training data [23]. A decision tree is then designed for every training subset until a tree ensemble has

been built. Every tree then inputs a unit vote influencing the outcomes of the incoming data instance cost label. Xu [24] diagnosed DM1 employing RS with a public hospital dataset, and this methodology performed better (85% success) than a number of other methodologies such as the ID3 (Iterative Dichotomiser 3) algorithm (78.57%), the naïve Bayes algorithm (79.89%), and the AdaBoost algorithm (84.19%).

SVM is a dual learning algorithm that undertakes example processing just by computing the dot-product. It is possible to efficiently compute such dot-products between feature vectors, employing a kernel function that does not obliterate every corresponding feature. Once the SVM learner has been supplied with the kernel function, it seeks out a hyperplane that can separate negative and positive examples and simultaneously maximize the size of their separation (margin). SVM is not prone to over-fitting and performs well in generalization as a result of the max-margin criterion employed throughout optimization. Although MLP solutions may solely be a local optimum, SVMs will always converge to a global optimum as a result of corresponding convex optimization formulations. The work [25] offered a useful method of hypoglycemic detection based on SVM, employing a galvanic skin response using skin temperatures, monitoring of heart rates, and a small band. Regrettably, the size and type of the dataset used in this research has unintentionally limited the applicability of the results. SVM revolves around employing high-dimensional feature spaces (built using transformational original variables) and the application of penalties to the resulting complexities by using a penalty term integrated within the error function [26]. Other approaches like the stacking-based General Regression Neural Network (GRNN) ensemble model are truly promising [27–29], but it has not been previously applied to DM1. In this sense, the authors of this work intend to analyze it in future research.

In relation to the foregoing, it can be concluded that these techniques have shown promise in terms of forecasting the dynamics of glycaemia. Nevertheless, as far as the authors are aware, no research has yet been undertaken employing selection techniques and a variety of forecasting algorithms utilizing real-world data and a proper feature set to reveal the most accurate of the options available.

3. Feature Selection and Forecasting Time Series

3.1. Feature Selection Techniques

Feature selection (FS) involves taking a particular dataset and selecting the most useful and applicable features from it. If we have a dataset with d input features, feature selection will create a set of k features in such a way that $k < d$, with k being the smallest possible collection of relevant and important features [30]. This means that the ML algorithm can be trained more quickly, the model becomes less complex and easier to decipher, forecasting power is improved, and overfitting is decreased through the selection of accurate feature arrangements, amongst other benefits.

Three types of feature selection methods are available [31], these being wrapper methods, filter methods, and embedded methods. Wrapper methods employ a combination of factors for decisions on the strength of forecasting. Standard wrapper techniques include Subset Selection, Forward Stepwise, and Backward Stepwise (RFE) [32]. The wrapper technique finds the optimal combination of features. This technique runs every variable past a test model created, using them for outcome assessment [33]. Of the three techniques, this demands the most computational power. Using the Subset selection technique, the model is fitted with all possible combinations of N features [34]. With Forward Stepwise techniques, we commence with a null model, i.e., one with only one variable, adding features singly and selecting the optimal model that scores the highest depending on the metrics used (f.i. A valuation of error). Having selected a predictor with this strategy, the model will never regress in the second stage. The process continues until we have the optimal feature subset, employing a stop criterion that sets the rules for the completion of the feature selection process. Conversely, Backward Stepwise Selection

(Recursive Feature Elimination) raises features as it processes. Since these techniques are not applied to every feature combination, they require significantly less computational power than straightforward subset selection [35]. Essentially, this technique is the opposite of the Forward Stepwise selection technique. The process begins with every predictor being present, erasing one after another and selecting the best model from the results. This requires essentially the same amount of computational power as Forward Selection. Some research has employed and made comparisons between Filter and Wrapper strategies [36].

Filter strategies are also referred to as Single Factor Analysis. Employing such techniques, an assessment is undertaken of each individual variable (feature)'s predictive power. A variety of statistical methods may be employed to ascertain how robust the predictions are [37]. One way of doing this is to undertake correlation of the features and the objectives (our predictions). The optimal features are those with the highest correlation.

The embedded Method (Shrinkage) represents a selection strategy using inbuilt variables. In this strategy features are neither selected nor excised. Certain controls are applied to parameter values (weights). Another technique is LASSO Regression. With this method, regularization is undertaken and certain regression coefficients tend towards zero [38]. As the coefficients fall towards zero they are dropped/rejected. Another technique is Ridge Regression (Tikhonov regularization), which incorporates a punishment increasing with the square of the coefficient greatness [39]. Every coefficient is diminished by an identical factor (which means no predictor undergoes elimination).

A selection of the above techniques will be employed in this research using a Ranker Strategy [40], which minimizes the metric Root Mean Squared Error (RMSE) and leads to reductions in the feature set. Both groups have differing approaches, one univariate and one multivariate. Univariate techniques are quicker and easier to scale, but they do not take account of variable dependencies. Conversely, multivariate techniques can model feature dependencies, but they are not as fast or as easy to scale as univariate techniques [41]. The methodology section will give further details on the selected techniques. If we minimize the metric, this leads to improvements in forecasting.

3.2. Forecasting

Once the FS is complete, we can begin the forecasting task in time series. Wolpert and Macready [42] stated that when we lack the information regarding the underlying model, we cannot say with certainty that any particular model will always outperform another. This means that the optimal strategy is to experiment with a number of techniques in order to discover the most effective model. This research has used both linear and non-linear techniques to focus on the algorithms that show the greatest promise.

Linear regression is one of the simplest techniques. In this model, we search for an estimate of the model parameters in order to minimize the sum of the squared errors [43]. The modifications incorporate partial least squares/penalized models, e.g., ridge regression or LASSO.

One advantage from the proponents of such models is that they are easy to interpret. Relationships are indicated by the coefficients and these are generally very simple to calculate, meaning that we can afford to employ a number of features. However, performance with these models may be limited [44]. Good results are achieved if the predictor/response relationship falls on a hyperplane. If we have higher-order relationships, e.g., like, cubic, or quadratic, such models may not accurately capture nonlinear relationships and thus we have to look for a different approach [45].

Certain models are capable of understanding non-linear trends. We do not need to know the precise type of nonlinearity prior to constructing our model. One of the most widely used models is Support Vector Machines (SVM). These are dual learning algorithms that compute the dot-products of data in processing [46]. Proper computation of such dot-products between variable rates may be achieved using a kernel function [47].

Using such a function, SVM learners seek out the hyperplane separating the examples with the maximum separation (margin) between them. It is recognized that SVMs are resistant to overfitting and perform well in terms of generalization as a result of the max-margin criterion employed during the optimization process. Additionally, although alternative solutions produce only local optimums, SVM will converge to a global optimum due to the corresponding convex optimization formulation [48].

There has been considerable interest in the Regression Trees family of modeling algorithms in recent times. Tree-based modeling employs if/then statements to find the predictors that will be used for data partitioning. In such subsets, a model is employed for forecasting outcomes [49]. Statistically, the addition of randomness when constructing the tree helps to reduce correlations between predictors. This is used in the Random Forests (RF) technique [50]. All models from the set are employed in building predictions for new datasets, and an average is taken of the predictions, which supplies the ultimate forecast.

RF models undertake variance reductions through the selection of robust complex learners with low bias levels. This decreases the number of errors and additionally proves strong in overcoming noisy responses [51].

Gaussian Processes (GPs) with Radial Basis Function Kernels (RBF) [52] and other forms of comparative strategy create consistency overall and allow for a limitless quantity of basic functions, but these are rarely used, even though some past research has demonstrated that it can show promise [53].

GP methodology is nonparametric, with a focus on taking discernible reactions from a variety of training data points (function values) and modeling them as multivariate standard random features [54]. It is assumed that there is a priority distribution of these function data values, guaranteeing that the function will operate smoothly.

When the comparing vectors are close (in the sensitivity and separation), the function values will be closely correlated, with decay occurring upon divergence. We may subsequently calculate how the unpredicted function data is distributed by employing an assumed distribution and applying a basic probability manipulation.

4. Database, Available Features and Target to Be Forecasted

4.1. Description of the Experiment

In order for this research to harvest a complete empirical collection of features, we created a new dataset by employing new ways of monitoring subjects. The glycaemia of 25 subjects underwent continuous monitoring for up to 14 days as they went about their normal routines. This monitoring campaign was awarded approval by the Ethical Research Commission of the University of Murcia on 25 January 2018 (Id.16 83/2017).

Each volunteer suffers DM1 being treated with a basal-bolus strategy, either employing slow insulin like Lantus, Levemir or Lantus—which creates a flat action curve—or fast insulin like Humalog-Lispro. Using slow insulin in basal coverage lasts for over 24 h; fast insulin is used to counter rises in glycaemia which may occur when eating or to counter hyperglycemia arising from other causes. Each subject signed an informed consent form prior to participation in the research.

There were 11 women and 14 men in the study cohort, each one receiving professional supervision and medical care. The monitoring undertaken was of a passive form that made no intervention in patient treatment, with every subject advised to continue with the instructions of their physicians. The subjects ranged from 18- to 56-year-olds, with an average age of 24.51 years, with the majority of subjects falling into the young adult category.

All patients had been suffering from diabetes for a minimum of five years; this time was set to ensure all patients had familiarity with the way the disease progresses.

All subjects were given complete information regarding the research aims. Generally, the subjects' condition was well controlled, all of them having a glycated hemoglobin(hbA1c) of between 6% and 7% when the experiment began.

Every patient claimed to be leading a healthy lifestyle, with none of them undertaking fewer than three sporting activities per week. Some scheduling was factored in in an attempt to make sure that every patient was following some form of routine without making drastic alterations to their daily routines. Each subject consumed a balanced diet that met their calorific needs. Subjects were asked to maintain their usual life habits and to continue following their endocrinologist's advice.

The patients in this research were given CGM sensors to wear, the model being the Freestyle Libre made by Abbott Company. This revolutionary device comprises a patch and a measuring device, permitting patients to make simple checks of their glycemic state (interstitial-glucose levels rather than blood-glucose levels). A notable feature of this device is that it allows the patient to check their glucose levels as frequently as they wish and collect data every 60 s. This device has been quite revolutionary as it is economically priced and achieves a reasonable level of accuracy (11.4% Mean Absolute Relative Difference, MARD). The subjects were asked to note down what fast-insulin dosages that had slow-insulin dosages, and also the carbohydrates they consumed through food, meaning that the data were empirical rather than subjective.

The CGM has a maximum lifespan of 14 days, but it can cease functioning prior to that. As it cannot be reattached once it has fallen off, it can cease working through accident, failed adhesion, or excessive humidity. In addition, the initial days of use can be inaccurate as the calibration is not solidified. It was proposed that data should be harvested from nine days of the usage period, with the initial days being excised as calibration was still taking place, and the final one is excised as the device may not have been able to cover the full 14-day lifespan. Thus, the experimental phase had 5400 h of data to consider.

Freestyle Libre is a device using flash glucose monitoring. Glucose levels are transmitted instantaneously when required, employing Near Field Communication (NFC) which needs the patient to actively require the data. Certain devices act as transductors NFC-Bluetooth (e.g., the popular miao-miao: https://miaomiao.cool/?lang=en, accessed January 2021). The Libre sensor can be attached to the device and this means that data can be transmitted to a smartphone at regular intervals.

The dataset was rounded out by use of the smart band Fitbit Charge HR©. This advanced fitness device automatically tracks various data and monitors the wearer's heart rate at all times. It can record sleep time, attitude climbed, step numbers, distance travelled, and heart rate. It connects using Bluetooth-low-energy and was linked with a computer and a small phone so that all necessary trends could be monitored. A number of other researchers have already employed Fitbit trackers to monitor the subjects' health [55]. All volunteers were given smart watches to keep a continuous record of their physical activity (step numbers) over the fortnight, along with heart rate and sleep data. Although these devices are not designed for precise medical use, it has been demonstrated that they are sufficiently accurate for the data to be used in research.

This monitoring was undertaken in 2018 and was continually supervised by the Endocrinology Departments of the Virgen de la Arrixaca and Morales Meseguer hospitals, two well-respected facilities in Murcia, Spain.

To the best of the authors' knowledge, unfortunately there is no previous work within the scientific literature with comprehensive data acquisition, since some previously published studies consider only partial monitoring, and collect only some of the features that can be recorded, or simply focus on a limited number of patients and/or data collection for just a few days.

Once all data had been acquired, preprocessing was undertaken on the dataset with outliers and gaps being cleaned. For cleaning outliers, extreme value analysis was employed, either by looking at scatter plots or by searching the values that were deviated more than double the mean. With gaps, interpolation methods were employed, with finger stick glucose values being added where possible. The sampling period was set at five minutes, sufficient to indicate tendencies and rapid changes but not so high that the ML algorithms would be overloaded. Data storage was undertaken in compliance with

the highest level of data protection regulations regarding personal information. Additionally, the Ethics Committee of the Universidad de Murcia, Spain, understood supervision of the way the patients were monitored.

The data gathered will be of great assistance to this research and other researchers in the future. Table 1 shows a variety of data relating to the population covered by our monitoring.

Table 1. Data regarding the patients considered in the study.

Population Feature	**Value**		
Subjects (Number)	**25**		
Sex	14 men–11 women		
Occupation	16 students–9 office workers		
Population Feature	Median	Min	Max
Age (years)	24.51	18	56
Body Mass Index (BMI, kg/m^2)	22.20	19.42	24.80
Duration of diabetes (years).	9	5	29
Fingersticks per day.	7	5	12
Insulin units per day (fast insulin + slow insulin, median).	47	36	59
HbA1C (%).	6.8	6.3	7.8

4.2. Available Features and Targets to Be Forecasted

Currently, the majority of past research into diabetes management systems has only considered insulin and glycemic levels, with some estimating the effect of meals; it would appear to be logical to add other variables that can affect glucose levels if they are susceptible to estimation or measurement. Overall, researchers have agreed that the remarkable variables should be meals, glycaemia, and insulin [56], which are the parameters used in most research. Some research recently has looked at other variables, chiefly exercise, both in vivo and in silico [57]. Additionally, they have also looked at temperature and heart rates. In this research we have harvested the following data every five minutes, with each element having some effect on a patient's glucose level:

- Glycaemia: A collection of previous measurements.
- Insulin injections: Previous values for fast insulin doses. For diabetic patients this hormone, generated exogenously, is the primary controller of how far blood glucose levels will fall.
- Meals: Previous values, as with insulin. It is noteworthy that the patient cohort all were experienced in counting their carbohydrates. All food by humans is converted and absorbed in the form of glucose, which is then released into the bloodstream, causing a virtually instantaneous rise in glycaemia.
- Exercise: Relevant historical data, with measurements in terms of steps taken. The muscles demand more glucose during physical activity; physical activity also enhances the circulation of the blood, making insulin more effective during exercise as the cellular barriers have greater permeability, meaning glucose has easier access to the cells.
- Heart rate: Contemporary and past values. The heart rate can be increased for a wide variety of reasons. Clearly it will rise during physical activity, but stress can be a contributor, as can hypo or hyperglycemia.
- Sleep: We collected data that only showed whether the subject was awake or asleep. It would appear logical to register sleep as being related to the length of sleep previously enjoyed. Poor quality nighttime sleep may cause insulin resistance and imbalances in glucose dynamics.

5. Methodology

5.1. The Waikato Environment for Knowledge Analysis (WEKA)

The University of Waikato, New Zealand, has produced the open source software Waikato Environment for Knowledge Analysis (WEKA v.3.8) (https://waikato.github.io/weka-wiki/, accessed December 2020). This free software is licensed with the GNU General Public License. WEKA comprises various algorithms and visualization tools to analyze data to use in predictive modeling, alongside graphical user interfaces allowing such functions to be easily accessed. A number of standard data mining routines can be performed with this software, particularly forecasting, modeling, feature selection, visualization, regression, classification, clustering, and data preprocessing.

The use of WEKA facilitates data entry, algorithm execution and visual context in the management of the entire process. This software has been successfully applied many times before, and is still being applied in recent literature. This way, Hussain et al. used WEKA in 2018 to study educational aspects with data mining techniques [58], and Kiranmai et al. also considered such software to classify electrical power problems [59]. WEKA is presently booming and new modules are developed every year, like those presented in 2009 by Lang et al. for deep learning [60].

WEKA includes specific libraries to tune the hyperparameters for each algorithm. In this sense, the Auto-WEKA package has been used [61]. Auto-WEKA considers the problem of simultaneously selecting a learning algorithm and setting its hyperparameters, overcoming the limitations of previous methods that address these issues in isolation. Auto-WEKA performs such task by using a fully automated approach, taking advantage of recent innovations in Bayesian optimization. Auto-WEKA helps to more effectively identify machine learning algorithms and hyperparameter settings, thereby achieving an improved performance.

5.2. Computer Hardware

Due to the computational demands of the ML algorithms considered in this work, they have been executed by a high-performance computer equipped with an AMD Ryzen 7 1700X processor, operating at 3.8 GHz with 32 GB DDR4 RAM at 2666 MHz CL19 and a Solid State Disk Samsung 970 Evo Plus M.2 1000 GB PCI-E 3.0.

5.3. Data Cleaning, Regularization and Lagged Variables

The database will undergo a transformation to provide the values that have been cleaned and a number of gaps that have been filled. Since the influence of certain features could be delayed, every feature apart from the date lacked 72 values (i.e., the previous six hours accounted for). WEKA's *TimeSeriesLagManager* permits lagged variables to be created when necessary.

5.4. Features Selection

Feature selection can be undertaken in WEKA with its intuitive graphical interface. The AttributeSelection module permits the specification of a variety of Attribute Evaluators and Search Methods. The testing of a number of combinations will be undertaken and an evaluation of them will be performed at the forecasting phase. Whichever features set provides the most accurate predictions will be selected.

5.4.1. Search Method

As mentioned above, a Ranker Strategy has been employed. This means of searching evaluates each feature one after the other and ranks them in order [62]. We can employ the identical name Ranker for the ordering within the AttributeSelection module.

5.4.2. Attribute Evaluators

Of the available feature selection methods in WEKA, the two most frequently employed will be chosen:

- Wrapper methods: employing the ClassifierAttributeEval routine within WEKA will permit evaluation or certain approaches. The predictors below will be executed.
    - Linear Regression: This allows for swift computation, with coefficients being fixed for all features.
    - Random Forest: As previously stated, this is a tree-based algorithm frequently employed for classification.
    - Multilayer Perceptron (MLP): This algorithm makes an estimation of the relative contributions of input units (which represent the attributes) and the output neurons (those which correspond with the problem classes) and uses the information to identify a subset of pertinent usable attributes to be employed in supervised pattern classification [63].
    - Instance-Based k-nearest neighbor algorithm (IBk) [64]: this is a K-nearest neighbor classifier that selects a suitable value for K on the basis of CV; it can also perform distance weighting.
- Filter Methods. For univariate methods, we will employ the predictors listed below.
    - Relief Attribute (Rlf) [65]: Relief feature selection works on the basis of creating a score by identifying feature value differences for nearest neighbor instance pairs.
    - Principal Component Analysis (PCA) [66]: With this method, we introduce a novel set of orthogonal coordinate axes, simultaneously maximizing sample data variants. This makes other directions with more minor variants have less significance and so they can be cleaned from the dataset. PCA is extremely effective in transforming data at lower dimensions and can also show us simplified underlying data patterns.

5.4.3. Generated Subsets

With a combination of exposed techniques as illustrated in Table 2, it is possible to generate seven subsets included in the original dataset without FS and subtests with reduced data. This can then be evaluated in the forecasting phase. For every exposed case of FS, the RMSE metric is the one that will undergo optimization. Additionally, we look at the prediction strategies that can be found in the subsection below from the original dataset. Table 3 is a tabulation of the various commands employed with WEKA, showing the parameters used.

Table 2. Applied Feature Selection techniques.

Search Method	Attribute Evaluator	Predictor	Acronym
Ranker	Wrapper (Classifier)	LR	Rnk-LR
		RF	Rnk-RF
		MLP	Rnk-MLP
		IBk	Rnk-IBk
	Filter	Relief	Rnk-Rlf
		PCA	Rnk-PCA

Rnk: Ranker; LR: Linear Regression; RF: Random Forest; MLP: Multi-Layer Perceptron; IBk: Instance-Based K-nearest neighbor; Rlf: Relief Attribute; PCA: Principal Component Analysis.

Table 3. WEKA commands for Feature Selection.

Technique	Command
Ranker	*weka.attributeSelection.Ranker -T -1.8E308 -N -1*
Classifier LR	*weka.attributeSelection.ClassifierAttributeEval -execution-slots 100 -B weka.classifiers.functions.LinearRegression -F 5 -T 0.01 -R 1 -E RMSE -- -S 0 -R 1.0E-8 -num-decimal-places 4" -S "weka.attributeSelection.Ranker -T -1.8E308 -N 100*
Classifier RF	*weka.attributeSelection.ClassifierAttributeEval -execution-slots 100 -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E RMSE -- -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1" -S "weka.attributeSelection.Ranker -T -1.8E308 -N 100*
Classifier MLP	*weka.attributeSelection.ClassifierAttributeEval -execution-slots 1 -B weka.classifiers.functions.MultilayerPerceptron -F 5 -T 0.01 -R 1 -E RMSE -- -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a" -S "weka.attributeSelection.Ranker -T -1. 1.8E308 -N 100"*
Classifier IBk	*weka.attributeSelection.ClassifierAttributeEval -execution-slots 1 –B weka.classifiers.lazy.IBk -F 5 -T 0.01 -R 1 -E RMSE -- -K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"" -S "weka.attributeSelection.Ranker -T -1.8E308 -N 100"*
Rlf	*"weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10" -S "weka.attributeSelection.Ranker -T -1.8E308 -N 100"*
PCA	*weka.attributeSelection.PrincipalComponents -R 0.95 -A 5" -S "weka.attributeSelection.Ranker -T 1.8E308 -N -1*

LR: Linear Regression; RF: Random Forest; MLP: Multi-Layer Perceptron; IBk: Instance-Based K-nearest neighbor; Rlf: Relief Attribute; PCA: Principal Component Analysis.

5.5. Data Modeling and Forecasting

Having obtained the seven reduced data subsets alongside the original dataset, we attempted to create predictions for future values in light of the previous time series collated for every dataset. An attempt has been made to forecast glycaemia for the subsequent 12 values. Since it is measured every five minutes, the maximum PH is fixed to 60 min. This allowed us to evaluate those predictions using real data. The training and test have been undertaken using Cross Validation for the time series [67,68]. This way, the data used in the training has been excluded from the training dataset.

To do this, we employed a WEKA (v.1.027) timeseriesForecasting module [69]. Therefore, the following algorithms will be used as crosses with every dataset, and also in relation to RMSE:

- Linear Regression (LR)
- Support Vector Machines (SVM)
- Random Forest (RF)
- Gaussian Process (GP)

Section 3.2 provides descriptions of all methods. Table 4 details the WEKA commands and each method's parameters.

Table 4. WEKA commands for forecasting.

Technique	Command
LR	*weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4*
RF	*weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1*
SVM	*weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1" -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007"*
GP	*weka.classifiers.functions.GaussianProcesses -L 1.0 -N 0 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -S 1*

LR: Linear Regression; RF: Random Forest; SVM: Support Vector Machines; GP: Gaussian Process.

6. Results and Discussion: Forecasting Performance

The different variable selection algorithms give rise to six reduced data subsets which, together with the original dataset, has led to seven cases on which to run the predictive techniques. Since the data of 25 patients were available, 150 processes of variable selection have been executed and together with the original dataset result in 175 data sets. The four predictive techniques have been applied to each of them, and a future prediction of the values for the following hour has been generated with five-minute intervals, that is, the prediction is made to the next 12 values, executing a CV. It should be noted that, as following the subject-wise scheme cannot be done, this could be a drawback of the study.

As an example of the first phase (training), from which a trained model is obtained, the graph in Figure 1 is obtained. It can be seen how the model generated with the subset generated after applying RF to patient '01' is used to predict at 60 min using RF as a predictive technique. As expected, the best inaccuracies correspond to periods of rapid oscillation and sharp variation in blood glucose, thus minimizing the error in hours without ups and downs.

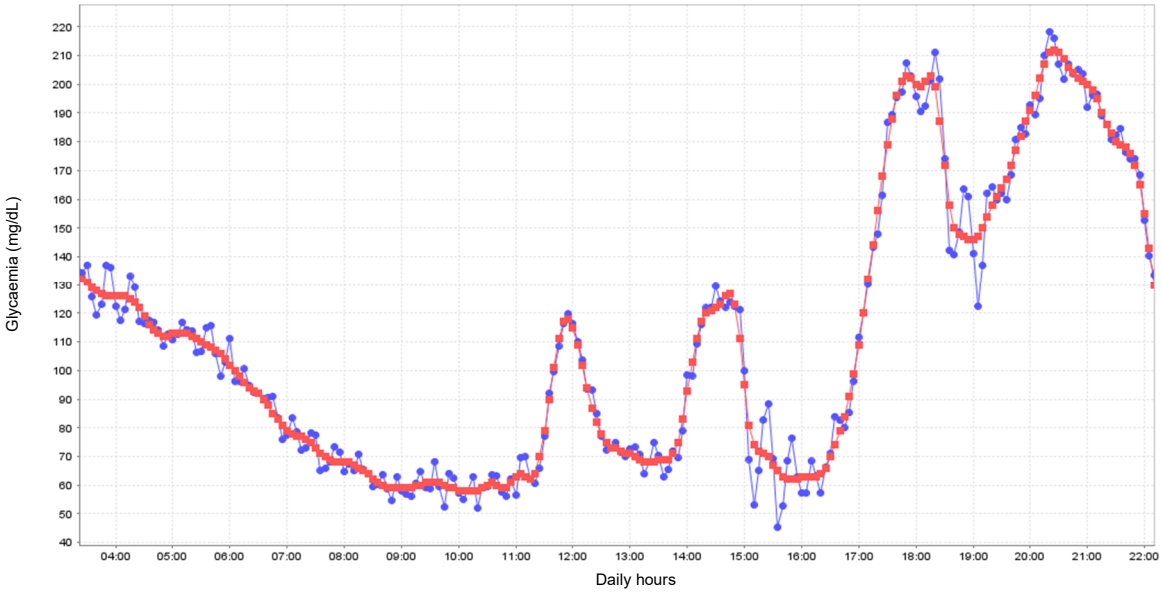

Figure 1. Example of training stage with RF algorithm of the subset RF from patient '01' for a 60 min PH glycemia prediction. Red, real data of glycaemia; Blue, modeled data.

The results of the forecasting task are tabulated in Table 5. With each predictive algorithm, and for each subset of data, we calculated the accuracy of the next 60 min (12 steps) of the glycaemia data. Using the CV technique, we obtained the RMSE (mg/dl) for each future step as an average of the 25 patients. Later, as a measure of performance, we further obtained an average of the 12 values of RMSE regarding each FS technique ($\overline{\text{RMSE}}$). We also estimated the standard deviation in each predicted series with the aim to infer the accuracy's variability. We performed the Shapiro–Wilk test to determine if the data presented a normal distribution for each 12-step prediction. The results showed that the data were normally distributed (p-values > 0.05).

Table 5. RMSE in test stage up to 12-steps glycemia forecasting. Averaged results of 25 subjects.

	RMSE (mg/dL)													
Subset FS	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	$\overline{\textbf{RMSE}}$	**Stnd Dev.**
Forecasting technique: LR														
No F.S.	9.32	15.15	18.94	22.72	26.23	29.30	31.26	33.33	35.48	37.69	40.03	42.55	28.50	10.32
LR	9.00	14.17	16.84	20.07	21.79	23.52	25.41	27.36	29.29	31.24	33.25	35.33	23.94	7.98
RF	9.29	14.53	17.16	20.26	21.82	23.38	25.08	26.84	28.58	30.32	32.09	33.91	23.60	7.41
MLP	9.22	14.44	17.09	20.25	21.89	23.51	25.26	27.06	28.83	30.61	32.42	34.29	23.74	7.57
IBk	9.51	14.88	17.57	20.69	22.26	23.83	25.54	27.31	29.07	30.84	32.63	34.48	24.05	7.50
Rlf	9.75	15.68	18.91	22.47	24.31	26.05	27.91	29.81	31.64	33.39	35.07	36.71	25.98	8.19
PCA	9.53	14.97	17.77	21.04	22.77	24.42	26.20	28.11	29.92	31.81	33.75	35.74	24.67	7.89
												$\overline{\overline{\textbf{RMSE}}}$	24.93	
Forecasting technique: RF														
No F.S.	13.17	20.87	24.78	28.65	31.03	31.78	32.40	32.92	33.33	33.66	33.92	34.15	29.22	6.50
LR	9.75	14.96	17.89	20.71	22.37	22.84	23.20	23.43	23.61	23.70	23.76	23.80	20.84	4.45
RF	7.91	13.21	16.22	19.08	19.74	20.21	20.57	20.83	21.01	21.16	21.26	21.33	18.54	4.14
MLP	8.88	14.18	17.14	19.97	21.68	22.18	22.53	22.78	22.94	23.04	23.10	23.10	20.13	4.52
IBk	7.95	13.29	16.27	19.12	20.81	21.30	21.87	22.15	22.35	22.50	22.58	22.59	19.40	4.63
Rlf	12.39	17.42	20.38	23.25	25.06	25.74	26.27	26.72	27.10	27.40	27.64	27.82	23.93	4.84
PCA	11.93	17.10	20.08	21.35	22.80	23.80	24.67	25.03	25.65	26.17	26.56	26.85	22.67	4.47
												$\overline{\overline{\textbf{RMSE}}}$	22.10	
Forecasting technique: SVM														
No F.S.	2.38	9.16	16.35	20.10	22.62	25.13	27.69	29.84	32.11	34.25	36.43	38.63	24.56	11.07
LR	1.99	7.64	13.53	16.45	18.37	20.39	22.51	24.27	26.11	27.85	29.65	31.48	20.02	8.96
RF	2.33	5.70	11.64	14.57	16.50	18.52	20.63	22.37	24.19	25.92	27.71	29.52	18.30	8.55
MLP	0.99	6.65	12.56	15.51	17.48	19.54	21.69	23.47	25.32	27.09	28.93	30.78	19.17	9.06
IBk	3.56	7.73	13.66	16.57	18.48	20.50	22.61	24.36	26.18	27.92	29.73	31.57	20.24	8.69
Rlf	4.02	8.44	14.82	18.02	20.14	22.30	24.40	25.98	27.62	29.12	30.65	32.16	21.47	8.82
PCA	3.26	7.77	13.74	16.67	18.62	20.64	22.76	24.52	26.35	28.09	29.89	31.71	20.33	8.79
												$\overline{\overline{\textbf{RMSE}}}$	20.58	
Forecasting technique: GP														
No F.S.	15.96	26.16	37.01	43.79	45.80	47.29	47.62	47.98	48.31	48.66	49.03	49.43	42.25	10.68
LR	12.08	25.82	31.17	33.37	34.34	34.79	35.02	35.15	35.23	35.28	35.32	35.35	31.91	6.83
RF	5.32	17.40	22.37	24.49	25.43	25.86	26.07	26.18	26.24	26.28	26.30	26.32	23.19	6.20
MLP	7.11	19.96	25.26	27.51	28.51	28.97	29.19	29.31	29.38	29.42	29.45	29.47	26.13	6.60
IBk	10.30	23.40	28.52	30.64	31.57	32.01	32.24	32.36	32.43	32.47	32.50	32.52	29.25	6.53
Rlf	7.84	24.73	32.56	36.34	38.26	39.28	39.84	40.17	40.38	40.51	40.60	40.66	35.10	9.78
PCA	15.62	24.24	27.85	29.37	30.02	30.31	30.44	30.50	30.53	30.54	30.55	30.55	28.38	4.42
												$\overline{\overline{\textbf{RMSE}}}$	30.89	

F.S.: Feature Selection; LR: Linear Regression; RF: Random Forest; SVM: Support Vector Machines; GP: Gaussian Process; MLP: Multi-Layer Perceptron; IBk: Instance-Based K-nearest neighbor; Rlf: Relief Attribute; PCA: Principal Component Analysis.

As stated in Table 5, we obtained the lower RMSE averaged between the 12 predictions ($\overline{\text{RMSE}}$) using RF as a foresight algorithm with the RF dataset ($\overline{\text{RMSE}}$ = 18.54 mg/dL) and in an average of the FS technique, the best performance is obtained using SVM as predictive technique ($\overline{\overline{\text{RMSE}}}$ = 20.58 mg/dL) but we have to note that the better performances are located on the early predictions (5, 10, 15 min), and then they rise. Figure 2 shows the evolution of the accuracy per forecasting algorithm with the different FS approaches.

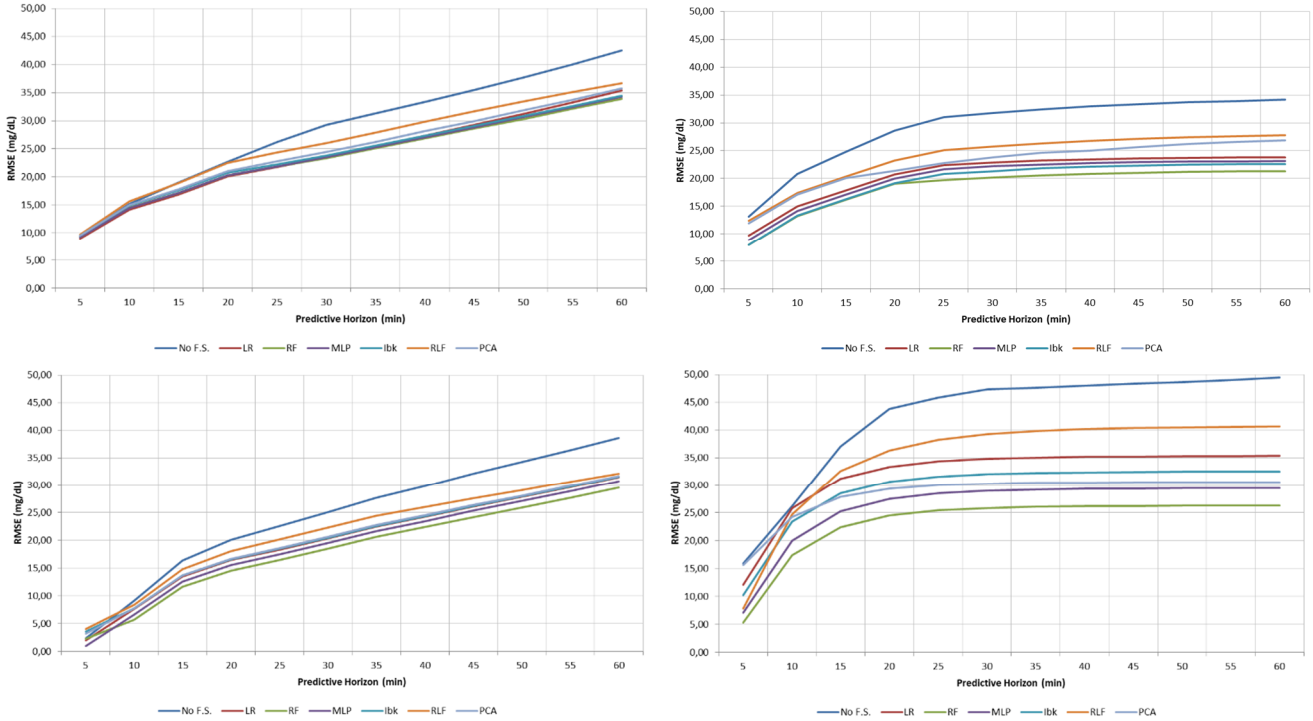

Figure 2. Test stage. RMSE up to 12-step glycemia forecasting. Averaged results of 25 subjects. **Upper left**, LR. **Upper right**, RF. **Lower left**, SVM. **Lower right**, GP. F.S.: Feature Selection; LR: Linear Regression; RF: Random Forest; SVM: Support Vector Machines; GP: Gaussian Process; MLP: Multi Layer Perceptron; IBk: Instance-Based K-nearest neighbor; Rlf: Relief Attribute; PCA: Principal Component Analysis.

In relation to Figure 2, we can observe how LR as a forecasting technique (upper left) offers an adequate behavior, but mainly in low PHs (5 or 10 min). Later, the error goes up for a further PH. It can be seen how the application of a variable selection method means a general improvement in performance, in light of the differences between the "no FS" line and the others.

This difference in performance according to FS technique is also observed in the RF predictive technique (above right). RF stands out as the best approach to select variables. In this case, RF as a forecasting technique generates good short-term accuracy, and in the long term it stabilizes at a value that may be acceptable under the best selective techniques.

SVM as a predictive algorithm (bottom left) presents an excellent performance at near PHs but in the long term it rises in a linear fashion. Again, it is observed how the selection of variables is necessary to reduce the error, but not all the techniques used result in a similar behavior. Lightweight RF as FS provides the best result.

GP shows the worst performance in terms of the prediction algorithm. The variability, according to the selection technique, is wide, the worst precision being when not using variable selection and the best when RF is used for this task.

In general, according to the mean values in Table 5, SVM is the best forecasting technique with an RMSE value of 20.58 mg/dL (considering all the FS techniques employed), and the best FS algorithm is RF. The latter (RF as the best FS technique) can be seen in the four predictive cases. It is also worth noting that the selection of variables always improves accuracy, and in some cases the technique itself can provide significant differences. Additionally, it is necessary to show that, for very short PHs, SVM works very well, but more broadly, RF could be a more balanced option. Variations in the error values at each step between the various 25 patients (standard deviation) are generally contained and do not present large variations between methods.

Figure 3 shows an example of how that prediction behaves at 60 min (12 steps). As can be seen, the blue line (prediction) follows the red line (actual data) quite closely, alt-

hough the error increases at situations of variation. This does not happen under situations of glycemic stability. Therefore, the patient's own control will influence the accuracy of the prediction. For this reason the predictions have been made with a CV method which has foreseen all types of real situations (stability and oscillations) throughout the days of monitoring of each subject.

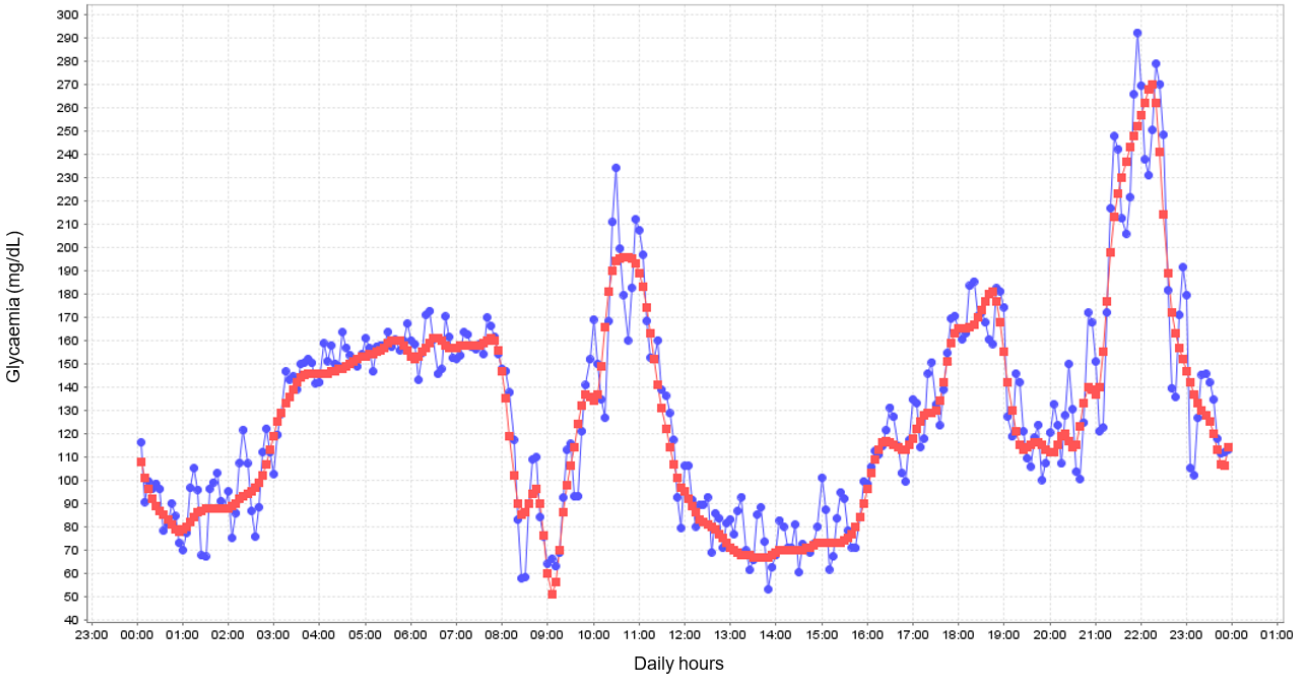

Figure 3. Example of forecasting stage with RF algorithm of the subset RF from patient '01' for a 60 min PH glycemia prediction. Red, real data of glycaemia; Blue, predicted data.

7. Conclusions and Future Works

In today's world, DM1 is a disease that requires multidisciplinary attention and research, not only from the medical field, but also from other disciplines such as data engineering. One of the first necessary phases before automating an artificial pancreas is to obtain a prediction of future glucose values in the most accurate possible way. Although some works have approached this aspect, a complete monitoring of the diabetic patient can provide more variables to refine the predictive process.

In this work, a monitoring time of 14 days was performed, including 25 DM1 patients, leading to an innovative dataset. Thus, it should be noted that the acquired dataset includes not only CGM estimations from a wide scope of people over a long timescale and in real life situations, but also incorporates other features such as insulin and eating times, and other related variables: heart rate, sleeping time and exercise.

During the glucose prediction process, it makes sense to establish an adequate preliminary phase of variable selection. Unfortunately, this significant phase has not always received the attention it deserves. This explains the reason why it is necessary to check to what extent the application of different feature selection techniques has influenced the accuracy of predictive algorithms. The results of the present work indicate that such facts are definitively relevant. Firstly, the precision obtained in the absence of any variable selection technique is clearly poorer than when incorporating it. In fact, among these techniques, the RF has been shown to be the best capable to refine the accuracy of glycemic prediction. In addition, numerous results have been obtained that indicate which forecasting algorithms clearly work better—that is, RF and SVM—and which FS techniques improve the performance of such techniques. In view of the obtained results, we can affirm that Random Forest (RF), as both a predictive algorithm and an FS strategy,

offers the best average performance (Root Median Square Error, RMSE = 18.54 mg/dL) throughout the 12 considered predictive horizons (up to 60 min in steps of 5 min), showing Support Vector Machines (SVM), the best accuracy as forecasting algorithm when considering, in turn, the average of the six FS techniques applied (RMSE = 20.58 mg/dL).

Future work will be focused on analyzing some promising previous approaches like the enhanced k-NN algorithm, which was successfully tested in [70] for biometric recognition, and the gradient boosting algorithm, using other databases like the D1NAMO project [71], which implied the monitoring of 20 healthy subjects and 9 patients by recording their electrocardiograms, breathing, accelerometer signals as well as glucose levels, including more biosensors that provide more variables in real time and thereby improving the accuracy of the glycaemia prediction and extending the PH within the glycemic series, and providing early warning of health monitoring [72].

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

Funding: Ignacio Rodríguez-Rodríguez would like to thank the support of Programa Operativo FEDER Andalucía 2014–2020 under Project No. UMA18-FEDERJA-023 and Universidad de Málaga, Campus de Excelencia Internacional Andalucía Tech.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethical Research Commission of the University of Murcia on 25 January 2018 (Id.16 83/2017).

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

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

References

1. Fowler, M.J. Diabetes: Magnitude and Mechanisms. *Clin. Diabetes* **2007**, *25*, 25–28, doi:10.2337/diaclin.25.1.25.
2. DeWitt, D.E.; Hirsch, I.B. Outpatient insulin therapy in type 1 and type 2 diabetes mellitus: Scientific review. *JAMA* **2003**, *289*, 2254–2264.
3. Davidson, M.B.; Davidson, M.B. *Diabetes Mellitus: Diagnosis and Treatment*; Saunders: Philadelphia, PA, USA, 1998.
4. Sherr, J.L.; Tauschmann, M.; Battelino, T.; De Bock, M.; Forlenza, G.; Roman, R.; Hood, K.; Maahs, D.M. ISPAD Clinical Practice Consensus Guidelines 2018: Diabetes technologies. *Pediatr. Diabetes* **2018**, *19*, 302–325, doi:10.1111/pedi.12731.
5. Westman, E.C.; Tondt, J.; Maguire, E.; Yancy, W.S., Jr. Implementing a low-carbohydrate, ketogenic diet to manage type 2 diabetes mellitus. *Expert Rev. Endocrinol. Metab.* **2018**, *13*, 263–272.
6. Kowalski, A. Can We Really Close the Loop and How Soon? Accelerating the Availability of an Artificial Pancreas: A Roadmap to Better Diabetes Outcomes. *Diabetes Technol. Ther.* **2009**, *11*, S113, doi:10.1089/dia.2009.0031.
7. Nguyen, B.P.; Ho, Y.; Wu, Z.; Chui, C.-K. Implementation of model predictive control with modified minimal model on low-power RISC microcontrollers. In Proceedings of the Third Symposium on Virtual Reality Modeling Language-VRML, Monterey, CA, USA, 16–19 February 2012, doi:10.1145/2350716.2350742.
8. Chui, C.-K.; Nguyen, B.P.; Ho, Y.; Wu, Z.; Nguyen, M.; Hong, G.S.; Mok, D.; Sun, S.; Chang, S. Embedded Real-Time Model Predictive Control for Glucose Regulation. In *XXVI Brazilian Congress on Biomedical Engineering*; Springer Nature: Berlin, Germany, 2013; Volume 39, pp. 1437–1440.
9. Eskaf, E.K.; Badawi, O.; Ritchings, T. Predicting blood glucose levels in diabetics using feature extraction and Artificial Neural Networks. In Proceedings of the 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, Syria, 7–11 April 2008; pp. 1–6.
10. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79, doi:10.1016/j.neucom.2017.11.077.
11. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324, doi:10.1016/s0004-3702(97)00043-x.

12. Balakrishnan, S.; Narayanaswamy, R.; Savarimuthu, N.; Samikannu, R. SVM ranking with backward search for feature selection in type II diabetes databases. In Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 12–15 October 2008; pp. 2628–2633.
13. Tomar, D.; Agarwal, S. Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes. *Adv. Artif. Neural Syst.* **2015**, *2015*, 1–10, doi:10.1155/2015/265637.
14. Rodríguez-Rodríguez, I.; Rodríguez, J.-V.; Zamora-Izquierdo, M. Variables to Be Monitored via Biomedical Sensors for Complete Type 1 Diabetes Mellitus Management: An Extension of the "On-Board" Concept. *J. Diabetes Res.* **2018**, *2018*, 1–14, doi:10.1155/2018/4826984.
15. Rodríguez-Rodríguez, I.; Rodríguez, J.-V.; González-Vidal, A.; Zamora, M.; Rodríguez, R.; Vidal, G. Feature Selection for Blood Glucose Level Prediction in Type 1 Diabetes Mellitus by Using the Sequential Input Selection Algorithm (SISAL). *Symmetry* **2019**, *11*, 1164, doi:10.3390/sym11091164.
16. Rodríguez-Rodríguez, I.; Chatzigiannakis, I.; Rodríguez, J.-V.; Maranghi, M.; Gentili, M.; Zamora-Izquierdo, M. Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through Machine Learning Techniques. *Sensors* **2019**, *19*, 4482, doi:10.3390/s19204482.
17. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Molina-García-Pardo, J.M.; Zamora-Izquierdo, M.Á.; Rodríguez-Rodríguez, M.T.M.I.I.; Martínez-Inglés, M.T. A Comparison of Different Models of Glycemia Dynamics for Improved Type 1 Diabetes Mellitus Management with Advanced Intelligent Analysis in an Internet of Things Context. *Appl. Sci.* **2020**, *10*, 4381.
18. Xie, J.; Wang, Q. Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison with Classical Time-Series Models. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 3101–3124, doi:10.1109/tbme.2020.2975959.
19. Sun, S.; Zhang, G.; Wang, C.; Zeng, W.; Li, J.; Grosse, R. Differentiable compositional kernel learning for Gaussian processes. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4828–4837.
20. Ortmann, L.; Shi, D.; Dassau, E.; Doyle, F.J.; Leonhardt, S.; Misgeld, B.J. Gaussian process-based model predictive control of blood glucose for patients with type 1 diabetes mellitus. In Proceedings of the 2017 11th Asian Control Conference (ASCC), Gold Coast, QLD, Australia, 17–20 December 2017.
21. Ortmann, L.; Shi, D.; Dassau, E.; Doyle, F.J.; Misgeld, B.J.; Leonhardt, S. Automated Insulin Delivery for Type 1 Diabetes Mellitus Patients using Gaussian Process-based Model Predictive Control. In Proceedings of the 2019 American Control Conference (ACC), Philadelphia, PA, USA, 10–12 July 2019.
22. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*, 1st ed.; The MIT Press: Cambridge, MA, USA, 2016; pp. 33–77.
23. Sage, A.J.; Genschel, U.; Nettleton, D. Tree aggregation for random forest class probability estimation. *Stat. Anal. Data Min.* **2020**, *13*, 134–150, doi:10.1002/sam.11446.
24. Xu, W.; Zhang, J.; Zhang, Q.; Wei, X. Risk prediction of type II diabetes based on random forest model. In Proceedings of the 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, India, 27–28 February 2017; pp. 382–386.
25. Marling, C.; Xia, L.; Bunescu, R.; Schwartz, F. Machine Learning Experiments with Noninvasive Sensors for Hypoglycemia Detection. In Proceedings of the IJCAI Workshop on Knowledge Discovery in Healthcare Data, New York, NY, USA, 19–24 June 2016.
26. Rodríguez-Rodríguez, I.; Zamora, M.Á.; Rodríguez, J.V. On predicting glycaemia in type 1 diabetes mellitus patients by using support vector machines. In Proceedings of the 1st International Conference on Internet of Things and Machine Learning, Liverpool, UK, 17–18 October 2017; pp. 1–2.
27. Izonin, I.; Tkachenko, R.; Verhun, V.; Zub, K. An approach towards missing data management using improved GRNN-SGTM ensemble method. *Eng. Sci. Technol. Int. J.* **2020**, in press, doi:10.1016/j.jestch.2020.10.005.
28. Tkachenko, R.; Izonin, I.; Kryvinska, N.; Dronyuk, I.; Zub, K. An Approach towards Increasing Prediction Accuracy for the Recovery of Missing IoT Data based on the GRNN-SGTM Ensemble. *Sensors* **2020**, *20*, 2625, doi:10.3390/s20092625.
29. Izonin, I.; Tkachenko, R.; Vitynskyi, P.; Zub, K.; Tkachenko, P.; Dronyuk, I. Stacking-based GRNN-SGTM Ensemble Model for Prediction Tasks. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Zallaq, Bahrain, 8–9 November 2020; pp. 326–330.
30. Guyon, I.; Elissee, A. An introduction to variable and feature selection. *J. Mach. Learn. Res*. **2003**, *3*, 1157–1182.
31. Sheikhpour, R.; Sarram, M.A.; Gharaghani, S.; Chahooki, M.A.Z. A Survey on semi-supervised feature selection methods. *Pattern Recognit.* **2017**, *64*, 141–158, doi:10.1016/j.patcog.2016.11.003.
32. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv* **2017**, arXiv:1707.08692.
33. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Pardo-Quiles, D.J.; Heras-González, P.; Chatzigiannakis, I. Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 8244.
34. Karegowda, A.G.; Manjunath, A.S.; Jayaram, M.A. Feature Subset Selection Problem using Wrapper Approach in Supervised Learning. *Int. J. Comput. Appl.* **2010**, *1*, 13–17, doi:10.5120/169–295.
35. Yang, K.; Yoon, H.; Shahabi, C. A supervised feature subset selection technique for multivariate time series. In Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning with Statistics, New Port Beach, CA, USA, 23 April 2005; pp. 92–101.

36. Crone, S.F.; Kourentzes, N. Feature selection for time series prediction—A combined filter and wrapper approach for neural networks. *Neurocomputing* **2010**, *73*, 1923–1936, doi:10.1016/j.neucom.2010.01.017.
37. Sánchez-Maroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter Methods for Feature Selection—A Comparative Study. In Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, Guilin, China, 30 October–1 November 2017; pp. 178–187.
38. Fonti, V.; Belitser, E. Feature Selection Using Lasso. *VU Amst. Res. Pap. Bus. Anal.* **2017**, *30*, 1–25.
39. Zhang, H.; Zhang, R.; Nie, F.; Li, X. A Generalized Uncorrelated Ridge Regression with Nonnegative Labels for Unsupervised Feature Selection. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2781–2785.
40. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2012**, *34*, 483–519, doi:10.1007/s10115-012-0487-8.
41. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Distributed feature selection: An application to microarray data classification. *Appl. Soft Comput.* **2015**, *30*, 136–150, doi:10.1016/j.asoc.2015.01.035.
42. Wolpert, D.; Macready, W. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82, doi:10.1109/4235.585893.
43. Shmueli, G.; Lichtendahl, K.C., Jr. *Practical Time Series Forecasting with r: A Hands-on Guide*; Axelrod Schnall Publishers: Green Cove Springs, FL, USA, 2016.
44. Faloutsos, C.; Gasthaus, J.; Januschowski, T.; Wang, Y. Forecasting big time series: Old and new. *Proc. VLDB Endow.* **2018**, *11*, 2102–2105.
45. Kalekar, P.S. *Time Series Forecasting Using Holt-Winters Exponential Smoothing*; Kanwal Rekhi School of Information Technology: Powai, Mumbai, 2004; pp. 1–13.
46. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
47. Schölkopf, B.; Smola, A.J. A short introduction to learning with kernels. In *Advanced Lectures on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 41–64.
48. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*, 1st ed.; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-6848-6.
49. Fierrez, J.; Morales, A.; Vera-Rodriguez, R.; Camacho, D. Multiple classifiers in biometrics. part 1: Fundamentals and review. *Inf. Fusion* **2018**, *44*, 57–64, doi:10.1016/j.inffus.2017.12.003.
50. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
51. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How Many Trees in A Random Forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 154–168.
52. Blomqvist, K.; Kaski, S.; Heinonen, M. Deep Convolutional Gaussian Processes. In Proceedings of the Mining Data for Financial Applications, Ghent, Belgium, 14–18 September 2020; pp. 582–597.
53. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Chatzigiannakis, I.; Zamora Izquierdo, M.Á. On the Possibility of Predicting Glycaemia 'On the Fly'with Constrained IoT Devices in Type 1 Diabetes Mellitus Patients. *Sensors* **2019**, *19*, 4538.
54. Seeger, M. Gaussian processes for machine learning. *Int. J. Neural Syst.* **2004**, *14*, 69–106.
55. Whelan, M.E.; Orme, M.; Kingsnorth, A.P.; Sherar, L.B.; Denton, F.L.; Esliger, D.W. Examining the Use of Glucose and Physical Activity Self-Monitoring Technologies in Individuals at Moderate to High Risk of Developing Type 2 Diabetes: Randomized Trial. *JMIR Mhealth Uhealth* **2019**, *7*, e14195, doi:10.2196/14195.
56. Bondia, J.; Vehi, J. Physiology-Based Interval Models: A Framework for Glucose Prediction Under Intra-patient Variability. In *Advances in Bioprocess Engineering and Technology*; Springer Nature: Berlin, Germany, 2015; pp. 159–181.
57. Garg, S.K.; Weinzimer, S.A.; Tamborlane, W.V.; Buckingham, B.A.; Bode, B.W.; Bailey, T.S.; Brazg, R.L.; Ilany, J.; Slover, R.H.; Anderson, S.M.; et al. Glucose Outcomes with the In-Home Use of a Hybrid Closed-Loop Insulin Delivery System in Adolescents and Adults with Type 1 Diabetes. *Diabetes Technol. Ther.* **2017**, *19*, 155–163, doi:10.1089/dia.2016.0421.
58. Hussain, S.; Dahan, N.A.; Ba-Alwi, F.M.; Ribata, N. Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indones. J. Electr. Eng. Comput. Sci.* **2018**, *9*, 447–459, doi:10.11591/ijeecs.v9.i2.pp447-459.
59. Kiranmai, S.A.; Laxmi, A.J. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. *Prot. Control. Mod. Power Syst.* **2018**, *3*, 29, doi:10.1186/s41601-018-0103-3.
60. Lang, S.; Bravo-Marquez, F.; Beckham, C.; Hall, M.; Frank, E. WekaDeeplearning4j: A deep learning package for Weka based on Deeplearning4j. *Knowl.-Based Syst.* **2019**, *178*, 48–50, doi:10.1016/j.knosys.2019.04.013.
61. Kotthoff, L.; Thornton, C.; Hoos, H.H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA: Automatic Model Selection and Hyperparameter Optimization in WEKA. In *Automated Machine Learning: Methods, Systems, Challenges*; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 81–95; ISBN 978-3-030-05318-5.
62. Novakovic, J.; Strbac, P.; Bulatovic, D. Toward optimal feature selection using ranking methods and classification algorithms. *Yugosl. J. Oper. Res.* **2011**, *21*, 119–135, doi:10.2298/yjor1101119n.
63. Gasca, E.; Sánchez, J.; Alonso, R. Eliminating redundancy and irrelevance using a new MLP-based feature selection method. *Pattern Recognit.* **2006**, *39*, 313–315, doi:10.1016/j.patcog.2005.09.002.
64. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66.
65. Kononenko, I. Estimating Attributes: Analysis and Extensions of RELIEF. In Proceedings of the European Conference on Machine Learning, Catania, Italy, 6–8 April 1994; pp. 171–182.
66. Abdi, H.; Williams, L.J. Principal component analysis. Wiley interdisciplinary reviews. *Comput. Stat.* **2010**, *2*, 433–459.

67. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213, doi:10.1016/j.ins.2011.12.028.
68. Snijders, T.A.B. On Cross-Validation for Predictor Evaluation in Time Series. In *Lecture Notes in Economics and Mathematical Systems*; Springer Nature: Berlin, Germany, 1988; Volume 307, pp. 56–69.
69. Frank, E.; Hall, M.A.; Holmes, G.; Kirkby, R.B.; Pfahringer, B.; Witten, I.H.; Trigg, L. Weka-A Machine Learning Workbench for Data Mining. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2009; pp. 1269–1277.
70. Nguyen, B.P.; Tay, W.-L.; Chui, C.-K. Robust Biometric Recognition from Palm Depth Images for Gloved Hands. *IEEE Trans. Hum.-Mach. Syst.* **2015**, *45*, 799–804, doi:10.1109/thms.2015.2453203.
71. Dubosson, F.; Ranvier, J.-E.; Bromuri, S.; Calbimonte, J.-P.; Ruiz, J.; Schumacher, M. The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Inform. Med. Unlocked* **2018**, *13*, 92–100, doi:10.1016/j.imu.2018.09.003.
72. Woo, W.L.; Koh, B.H.; Gao, B.; Nwoye, E.O.; Wei, B.; Dlay, S.S. Early Warning of Health Condition and Visual Analytics for Multivariable Vital Signs. In Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things, Sanya, China, 24–26 April 2020; pp. 206–211.