




Article

# A Method to Detect Type 1 Diabetes Based on Physical Activity Measurements Using a Mobile Device

Anna Czmil <sup>\*</sup>, Sylwester Czmil  and Damian Mazur 

Faculty of Electrical and Computer Engineering, Rzeszow University of Technology, 35-959 Rzeszow, Poland; sylwesterczmil@gmail.com (S.C.); mazur@prz.edu.pl (D.M.)

\* Correspondence: czmilanna@gmail.com

Received: 28 May 2019; Accepted: 19 June 2019; Published: 22 June 2019



**Featured Application:** Non-invasive method of type 1 diabetes detection based on physical activity measurement.

**Abstract:** Type 1 diabetes is a chronic disease marked by high blood glucose levels, called hyperglycemia. Diagnosis of diabetes typically requires one or more blood tests. The aim of this paper is to discuss a non-invasive method of type 1 diabetes detection, based on physical activity measurement. We solved a binary classification problem using a variety of computational intelligence methods, including non-linear classification algorithms, which were applied and comparatively assessed. Prediction of disease presence among children and adolescents was evaluated using performance measures, such as accuracy, sensitivity, specificity, precision, the goodness index, and AUC. The most satisfying results were obtained when using the random forest method. The primary parameters in disease detection were weekly step count and the weekly number of vigorous activity minutes. The dependance between the weekly number of steps and the type 1 diabetes presence was established after an insightful analysis of data using classification and clustering algorithms. The findings have shown promising results that type 1 diabetes can be diagnosed using physical activity measurement. This is essential regarding the non-invasiveness and flexibility of the detection method, which can be tested at any time anywhere. The proposed technique can be implemented on a mobile device.

**Keywords:** type 1 diabetes; classification; physical activity; artificial intelligence

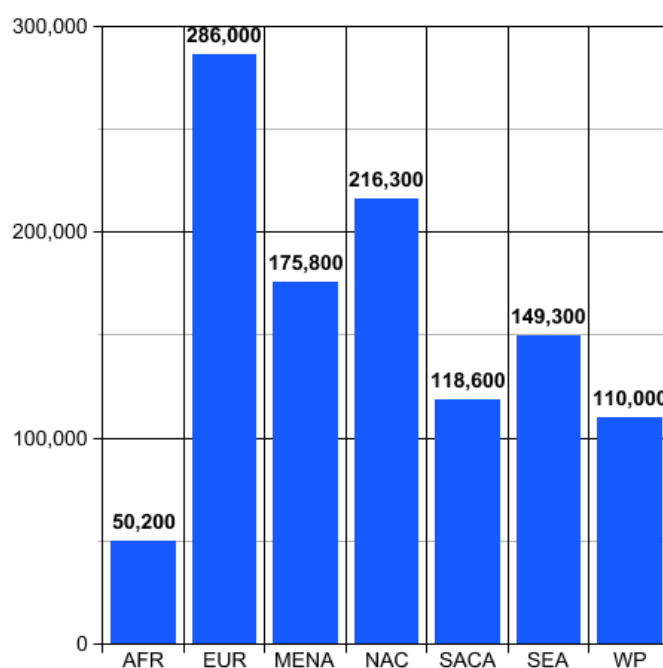
## 1. Introduction

Diabetes mellitus is a group of metabolic diseases that is characterized by hyperglycemia and results from defects in insulin action, insulin secretion, or both [1]. Elevated blood glucose connected with this disease can cause dysfunction and failure of various organs, which are the effects of long-term diabetes. Currently, according to the WHO and American Diabetes Association classification (ADA), there are four types of diabetes: type 1, type 2, other specific types of diabetes, and gestational diabetes [2,3].

Type 1 diabetes causes the patient's blood glucose to become too high. This happens when his or her body cannot produce enough insulin, which controls blood glucose. Patients need daily injections of insulin to keep blood glucose levels under control. It is one of the leading health problems in Poland and Europe, for people of all ages. It causes constant damage to health and contributes to premature death [4,5]. According to the International Diabetes Federation estimation, the incidence of type 1 diabetes among children and adolescents under the age of 15 years is increasing in many countries,

and the overall annual increase is estimated to be around 3%, with strong indications of geographic differences. More than 96,000 children and adolescents under 15 are estimated to be diagnosed with type 1 diabetes annually. The number is estimated to be more than 132,600 when the age range is extended to 20 years. In total, more than one million children globally and adolescents below 20 are estimated to have type 1 diabetes [6].

There are large regional differences in the number of children and adolescents with type 1 diabetes. Last year, in Europe, there were 28.4% of children and adolescents with type 1 diabetes and 21.5% in North America and the Caribbean. The United States, India, and Brazil have the largest incidence and prevalence of children with type 1 diabetes under both age groups below 15 and 20 years old (Figure 1) [6].



**Figure 1.** Estimated number of children and adolescents <20 years with type 1 diabetes by IDF region, 2017 [6].

Type 1 diabetes is described as the most prevalent metabolic disease and the third most common and irreversible chronic disease in childhood, especially below 15 years of age [7]. Despite great progress in medicine, diabetes is an incurable disease, and it is an extraordinary burden on patients and their families. Due to its chronic, progressive, and incurable nature, it greatly affects adolescents, in particular basically their self-esteem, educational opportunities, and lifestyle. Children and adolescents with type 1 diabetes must face many problems related to treatment restrictions.

Measurement of blood sugar is the basic test most often ordered by doctors to detect carbohydrate tolerance disorders and also to diagnose and monitor the treatment of diabetes. Blood is drawn for testing on an empty stomach, followed by a meal or after administration of glucose solution. Serious barriers in the treatment of diabetes among children are problems with painful injections or blood tests, shame about diabetes, arguing with parents about the plan for diabetes control, and compliance. Particularly troublesome are activities related to measuring the level of glucose in the blood, making injections of insulin, exercising, controlling the content of carbohydrate dietary exchanges in the diet, wearing a diabetic or information bracelet, carrying sweets for hypoglycemia, and eating snacks [8].

An additional problem is the fact that the symptoms of diabetes are often ambiguous. They may be confused or attributed to other diseases. Diabetes can only be unequivocally diagnosed when a glucose load test is performed. Too late of a diagnosis of diabetes in childhood can lead to serious

changes, such as destruction of blood vessels, visual disturbances, and problems with the nervous system and kidneys. Very serious diabetes, having been unrecognized for a long time, may endanger children's lives; therefore, extraordinary vigilance should be maintained while observing children, in order to react in time to the first signals of the disease [9,10].

While analyzing the information above, the question arises whether it is possible to diagnose diabetes without performing blood tests. The present work aims to diagnose type 1 diabetes among children based on their physical activity. Selected classification algorithms are compared to obtain the most satisfying results. The promising results encourage developing an application using computational intelligence methods.

## 2. Background

### 2.1. Available Methods of Assessing Physical Activity

Physical activity results in an increase in energy expenditure above resting levels. The rate of energy expenditure is directly linked to the intensity of the activity [11]. Physical activity can be classified according to the Borg scale, ranging from sedentary, light, moderate, to vigorous activities [12].

Currently, there are many methods that allow determining the parameters of physical activity with high accuracy. These include all monitors like pedometers and accelerometers that have motion sensors and are worn on the body of the subject to perform various motion measurements, e.g., step count, the duration of physical activity, and its intensity [13].

### 2.2. Pedometers and Accelerometers in Physical Activity Measuring

The simplest and most popular devices allowing activity measurements are pedometers, which record the number of steps. Thanks to the ability to display the result on a regular basis, they are considered as a motivating tool to perform more physical activity in everyday life. However, measurements by pedometers in scientific research have many limitations. Devices provide information on the frequency of movement, but they do not determine the intensity of physical exercise. Pedometer step counts are also more inaccurate at slow speeds (<60 m/min); therefore, they may be inappropriate for older adults, and the result may not be reliable. Pedometer readings can also vary according to where the pedometer is mounted. In addition, its weakness is also the possibility of falsifying and increasing results by intentionally shaking the device or by shocks caused by driving a car, which do not prove that the subject was more active [14].

Currently, the most accurate motion sensors used to assess physical activity are accelerometers. The devices detect the acceleration of body movement, giving the opportunity to measure reliably the intensity and duration of physical activity, as well as the number of steps taken, and sedentary analysis [15]. Those parameters of the motion are read by the piezoelectric sensor, which converts the analog signal into the digital one in the range (0.1–3.6 Hz). Thanks to this, very accurate monitoring of physical activity is possible. An example of a commonly-used accelerometer is ActiGraph.

### 2.3. ActiGraph Activity Monitor

ActiGraph has been used in large-scale field studies and has become the de facto standard device for objective physical activity monitoring [16]. It is particularly recommended for examination of children and adolescents because it allows for detection of acceleration in three planes of motion, which provides more accurate analysis of the movement relative to pedometers. This is especially important in the case of children's examination because the device records all forms of physical activity, such as doing push-ups or climbing. Many publications describe the advantages of using accelerometers in scientific research, such as objectivity, non-invasiveness, and accuracy, while maintaining the comfort of the user [15].

Published findings related to the application of ActiGraph concerned with exploring differences in daily physical activity profiles among individuals with mild Alzheimer's disease were compared to a control group [17]. Features that can be derived from the accelerometer have been also used to recognize the presence and severity of motor fluctuations in patients with Parkinson's disease [18]. It has been also used with measurements of physical activity to evaluate the effectiveness of surgical and therapy-based interventions in children with cerebral palsy or to derive diurnal rest-activity patterns from actigraphy in adolescents and to analyze associations with adiposity measures and cardiometabolic risk factors [19,20].

However, ActiGraph activity monitors have limited memory and battery capacity to store raw signal data and are additionally quite expensive. One of the current models, ActiGraph wGT3X-BT, currently sells for 225 USD [21]. The costs of devices may vary if bought separately, as compared to bulk orders.

Due to memory limitations, information about movement is read by the accelerometer in the form of the number of pulses (named counts), which are added up in the designated time unit [22]. A count is a unit aimed to be proportional to the average overall acceleration of the human body in a specified period of time. The sum of the received counts is converted into the intensity of physical activity, categorized as sedentary behaviors, light physical activity (LPA), moderate PA (MPA), and vigorous PA (VPA).

There are commonly-used regression equations named as cut points for the ActiGraph accelerometers in predicting energy expenditure (EE) in children and adolescents [23]. The cut points are derived as a part of published research aimed at quantifying activity levels using ActiGraph products. All cut point sets are scaled to 60-s epochs.

In this study, the parameters of physical activity are calculated according to the Freedson Children (2005) model. Definitions of the cut point levels for this model are given in Table 1.

**Table 1.** Freedson Children ActiGraph cut points.

Activity Label	Cut Point	
	From	To
Sedentary	0	149
Light	150	499
Moderate	500	3999
Vigorous	4000	7599
Very Vigorous	7600	$\infty$

#### 2.4. Methods to Compare New and Traditional Accelerometer Data

There are many publications describing how to convert a raw accelerometer signal into the output data of the ActiGraph [16,24,25]. Such data can be obtained using a common smartphone, which is equipped with an accelerometer and a pedometer. Mapping the conversion of counts will allow performing tests in an inexpensive and easy way, which will be comparable to those obtained using the ActiGraph activity monitor.

The research literature describes that counts are calculated as the area under the filtered and rectified (non-negative) curve. The ratio between raw acceleration signal and counts is likely to be brand specific [16]. The experiment described in the literature showed that a third-order Butterworth filter resulted in the highest correlation between ActiGraph counts and unscaled raw accelerometer counts ( $r = 0.975$ ,  $p < 0.01$ ) [24].

The complete method of the conversion of raw accelerometer data to the output the ActiGraph signal is presented below as steps. First, it is necessary to gather 60 s of analog accelerometer reads and calculate the Euclidean distance on analog data in order to create one signal from three axes. Second, this signal should be processed using a third-order Butterworth filter. Next, the area under the filtered

and rectified signal should be calculated. Then, the result should be labeled by type of activity (i.e., sedentary, vigorous, etc.) using predefined cut points and a count of the selected incremented label. All steps should be repeated until enough data are collected.

This method allows for consistency with traditional physical activity measurements so that it is possible to make a historical analysis and comparisons.

### 3. Materials and Methods

#### 3.1. Data Source

The dataset was collected from a group of schoolchildren between the ages of 6 and 18 being under the care of the diabetic clinic for children at Rzeszow State Hospital in Poland in 2016 by E.Czeczek-Lewandowska as a part of her Ph.D. thesis research [8]. The dataset was divided into two groups based on the results of HbA1c glycated hemoglobin tests for diabetes that were read from the patient's medical records provided by the diabetic clinic with parental consent. The analysis included the last two results from the maximum period of one year prior to the study, on the basis of which the arithmetic mean was calculated.

Of the 451 children that took part in the research, the inclusion and exclusion criteria were extracted and analyzed. The eligibility criteria that were applied were: ages between 6 and 18, type 1 diabetes diagnosed a minimum of one year prior to the examination, HbA1c values determined at least twice in the year prior to the start of the study, informing parents about the study and child consent, required physical activity record length (excluding night hours and activities performed in contact with water), and training the parent and child in terms of using the accelerometer. Children who did not meet the inclusion criteria, were diagnosed with type 2 diabetes or other metabolic disorders, had current complications in the course of diabetes, and became sick during the study period were excluded from the study. Additional excluding criteria were exceptionally bad weather conditions, a period of holidays, and holiday break during the study period. Finally, the study group consisted of 215 children with type 1 diabetes and 115 healthy children from a control group. Nine parameters for each child were collected and are listed below.

1. General and BMI parameters:
  - Age
  - Sex
  - Weight
  - Height
2. Physical activity parameters (per week):
  - Step count
  - Sedentary activity minutes
  - Light activity minutes
  - Moderate activity minutes
  - Vigorous activity minutes
3. Type 1 diabetes presence (binary parameter)

The weight and height of the body were obtained using a Radwag WPT 60/150 OW electronic scale during a three-stage measurement. The level of physical activity was assessed with a hip-worn ActiGraph wGT3X-BT activity monitor used by the children 12 h a day for a week, excluding night time and activities performed in contact with water, i.e., bath, swimming pool. The parameters of physical activity were calculated according to the Freedson Children (2005) method.

### 3.2. Classification Methods

Classification systems have an important role in decision-making tasks by categorizing the available information based on some criteria [26]. The purpose of this research was to assess the relative efficacy of some well-known classification methods. We have considered classification techniques that are based on statistical and AI techniques. A brief review of the relevant classification methods is presented in this section.

#### 3.2.1. Support Vector Machine

Support vector machine (SVM) is a classification algorithm used for finding an optimal hyperplane that maximizes the margin between classes. That hyperplane is orientated in such a way that it is as far as possible from the closest data points from each of the classes. These closest points are called support vectors [27]. The key element of the SVM algorithm is the kernel function. It transforms a non-linear feature space into a linear one before the hyperplane search [28].

#### 3.2.2. Probabilistic Neural Network

The probabilistic neural network (PNN) is a feedforward neural network model. It consists of input, pattern, summation, and output layers. The input layer is represented by the features of the input vector. The pattern layer is composed of as many neurons as learning samples. The summation layer consists of  $N$  neurons where each of them computes the signal only for patterns that belong to the  $n^{\text{th}}$  class. The output layer is used to yield the decision; its result with the largest probability value is 1, and the rest of the outputs are 0 [29].

#### 3.2.3. Multilayer Perceptron

Multilayer perceptron (MLP) is a feedforward artificial neural network that uses the backpropagation technique for training. It is composed of one or more layers of neurons. Data are transferred to the input layer; there may be one or more hidden layers; and predictions are made on the output layer [30].

#### 3.2.4. Group Method of Data Handling

The group method of data handling (GMDH) is a family of inductive algorithms of multi-parametric datasets. It features the fully-automatic parametric and structural optimization of models. GMDH is used for constructing a high-order regression-type polynomial [31].

#### 3.2.5. Gene Expression Programming

Gene expression programming (GEP) is an evolutionary algorithm that creates models, equations, or computer programs. GEP programs are encoded in the so-called chromosomes, which are mutated by computing the expression of each chromosome. Next, the predefined genetic operators are applied, and the fitness is calculated. Finally, the best chromosomes are selected to reproduce [32].

#### 3.2.6. Linear Regression

Linear regression is one of the simplest and best known algorithms in statistics and machine learning used for finding a linear relationship between the target and one or more predictors. The core idea of linear regression is to obtain a line that best fits the data [33].

#### 3.2.7. Radial Basis Function Network

The radial basis function network (RBF) is an artificial neural network that uses radial basis functions as activation functions. The output of the RBF network is composed of neuron parameters and radial basis functions of the inputs [34].

### 3.2.8. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset with one or more independent variables that determine an outcome. The goal of logistic regression is to find the best fitting model to describe the relationship between the binary dependent variable and a set of independent variables [35].

### 3.2.9. Decision Tree

The decision tree (DT) is a type of model used for both classification and regression. Trees answer sequential questions, which are sent down a certain route of the tree given the answer. They are intuitive and provide one of the simplest portrayals for classification purposes. Tree depth represents how many questions are asked before reaching the predicted classification [36].

### 3.2.10. Random Forests

Random forests (RF) are a classification algorithm that is a combination of decision tree predictors so that each of them depends on the values of a randomly -elected independent vector with the same distribution for all trees in the forest [37]. After training, predictions for unseen samples can be made by taking the majority vote [36].

## 3.3. Validation Methods

Commonly-used evaluation measures are precision, sensitivity, and accuracy. These measures can be defined with the help of four cardinalities of the confusion matrix, namely the truth positive (TP), true negative (TN), false positive (FP), and false negative (FN) [38].

### 3.3.1. Accuracy

The accuracy metric measures the total number of correct classifications (true positives and true negatives) [38].

$$ACC_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}, TP_i + TN_i + FP_i + FN_i > 0 \quad (1)$$

### 3.3.2. Sensitivity

The sensitivity (recall) measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of children with type 1 diabetes who are correctly identified as having the condition): [38].

$$SE_i = \frac{TP_i}{TP_i + FN_i}, TP_i + FN_i > 0 \quad (2)$$

### 3.3.3. Specificity

The specificity measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy children who are correctly identified as not having the condition): [38].

$$SP_i = \frac{TN_i}{TN_i + FP_i}, TN_i + FP_i > 0 \quad (3)$$

### 3.3.4. Precision

The precision metrics determine the quality of positive predictions (true positives and false positives): [38].

$$PPV_i = \frac{TP_i}{TP_i + FP_i}, TP_i + FP_i > 0 \quad (4)$$

### 3.3.5. AUC

For a binary classification problem, the evaluation of the performance is typically illustrated with the receiver operating characteristic (ROC) curve, which plots the true positives versus the false positive rate at various threshold settings. It is convenient to reduce it to a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve (AUC). An ideal classifier achieves an AUC equal to 1, while the classifier that makes a random decision achieves an AUC equal to 0.5 [38,39].

### 3.3.6. Goodness Index

The goodness index (G) represents the Euclidean distance between the evaluated point in the receiver operating characteristic space and the point (0,1), which represents the perfect classifier that classifies all positive cases and negative cases correctly.

$$G_i = \sqrt{\left(1 - \left(\frac{TP_i}{TP_i + FN_i}\right)\right)^2 + \left(1 - \left(\frac{TN_i}{FP_i + TN_i}\right)\right)^2} \tag{5}$$

G can assume values between 0 and  $\sqrt{2}$ , and a classifier can be considered as:

- optimum, when  $G \leq 0.25$ ,
- good, when  $0.25 < G < 0.70$ ,
- random, if  $G = 0.70$ ,
- bad, if  $G > 0.70$  [40].

The G value result analysis allows evaluating the best-performing classifier [28].

## 3.4. Other Data Analysis Methods

### 3.4.1. Clustering Method

The *k*-means clustering algorithm is one of the most popular clustering algorithms, which is used to find groups that are not explicitly labeled in the data. It uses iterative refinement to produce a final result. The inputs of the algorithm are the dataset and the number of clusters *k*. A cluster is a collection of data points that have been aggregated together because of certain similarities, and the dataset is a collection of features for each data point. The algorithms start with initial estimates for the *k* centroids, which can either be randomly initialized or randomly selected from the dataset. Then, the algorithm iterates between two steps:

- Data assignment: each data point is assigned to its nearest centroid, based on the squared Euclidean distance. If  $c_i$  is the collection of centroids in set *C*, then each data point *x* is assigned to a cluster based on:

$$\arg \min_{c_i \in C} dist(c_i, x)^2 \tag{6}$$

where  $dist(\cdot)$  is the standard Euclidean distance.  $S_i$  is the set of data point assignments for each *i*<sup>th</sup> cluster centroid.

- Centroid update: centroids are recomputed by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \tag{7}$$

The algorithm iterates between those two steps until convergence. Convergence is reached when the computed centroids do not change or the centroids and the assigned points oscillate back and forth from one iteration to the next one. The result may be a local optimum, so assessing more than one run of the algorithm with randomized starting centroids may give a better outcome [41].



### 3.4.2. Feature Selection Methods

Feature selection is the first and fundamental step in data analysis. This is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection methods aid in creating an accurate predictive model by choosing only features that are relevant. Irrelevant features in the dataset can decrease the performance of the models; redundant data can allow a greater opportunity to make decisions based on noise and increase algorithm complexity, while algorithms are trained more slowly.

There are three general classes of feature selection algorithms: filter methods, wrapper methods, and embedded methods. Filter feature selection methods apply a statistical measure to assign a score to each feature. The features are ranked by the score and are either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable. Examples of some filter methods include correlation coefficient scores and information gain. These methods are used to create the feature ranking [42].

Pearson's correlation coefficient is one of the methods of measuring the association between variables of interest, and it is based on the covariance method. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship [43].

Entropy measures the amount of uncertainty in the dataset. The information gain is based on the decrease in entropy after splitting a dataset on an attribute [44]. It is used to generate a decision tree from a dataset. Constructing a decision tree comes down to finding an attribute that returns the highest information gain.

The information gain  $IG$  is the change in information entropy  $H$  from a prior state to a state that takes some information as given:

$$IG(d | a) = H(d) - H(d | a) \quad (8)$$

where  $H(d | a)$  is the conditional entropy of decision  $d$  given attribute  $a$  and  $H(d)$  is the entropy of decision  $d$ , which is equal to:

$$H(d) = - \sum_{i=1}^k p(d_i) \cdot \ln p(d_i) \quad (9)$$

Information gain can be calculated for each remaining attribute. The attribute with the largest information gain is used to split the dataset on this iteration [45]:

$$IG(d | a) = H(d) - \sum_{j=1}^l p(a_j) \cdot H(d | a_j) \quad (10)$$

## 4. Results

### 4.1. Data Analysis Results

The aim of this research was to answer whether type 1 diabetes among children and adolescents can be diagnosed based on physical activity. We defined the prediction problem of type 1 diabetes presence among children as a binary classification problem. The results were obtained using DTREG, Weka, and Python Scikit-Learn software packages [46–48].

The assessment of physical activity impact on the prevalence of type 1 diabetes among children and adolescents was based on parameters closely related to the intensity of physical activity. These parameters were calculated according to the Freedson Children (2005) model (Table 1). We considered the dataset consisting of parameter values of 215 sick and 115 healthy children. The selected classification parameters set was composed of the total number of steps and sedentary, light, moderate, and vigorous activity minutes per week.

Subsequently, we decided to create a feature ranking (FR) automatically. FR specifies the significance of features for a problem by ranking features according to their importance in the model using ranking algorithms [42]. The FR based on correlation coefficient scores was performed, and the results are presented in Table 2.

**Table 2.** Correlation coefficient feature ranking.

	Feature Name	Score
1	step count	0.2362
2	vigorous activity minutes	0.0505
3	moderate activity minutes	0.0469
4	sedentary activity minutes	0.0408
5	light activity minutes	0.0127

Due to the fact that evaluating the entropy is a key step in the decision tree algorithm, it was used to calculate the homogeneity of a sample, and we decided to create FR based on information gain, which is based on the entropy. The results are presented in Table 3.

**Table 3.** Information gain feature ranking.

	Feature Name	Score
1	vigorous activity minutes	0.1435
2	moderate activity minutes	0.1375
3	step count	0.084
...	...	0

The data presented in Table 2 showed that the most significant parameter was the step count (per week). Data presented in Table 3 resulted in three important parameters, i.e., vigorous activity minutes, moderate activity minutes, and step count.

The presented results of the FR are purely illustrative, because threshold values were set to exclude unimportant parameters. We decided to use the classification of all physical activity parameters.

#### 4.2. Classification Result

Firstly, we built a decision tree with an overall goal to extract general information from a dataset and transform that information into a structure that can be understood by an ordinary user. A decision tree was built from physical activity parameters, i.e., the total number of steps and the groups of sedentary, light, moderate and vigorous activities, using the implementation of the c4.5 algorithm, called J48, from the Weka software package. The algorithm was started with default values, such as the confidence threshold for pruning the set to 0.25 and a minimum number of instances per leaf equal to two.

At each node of the tree, the algorithm chose the attribute of the data that most effectively split the set of samples into subsets enriched in one class or the other. The splitting criterion was the normalized information gain. The information gain feature ranking results described in Section 4.1 had the vigorous activity minutes parameter in the first place. Hence, it could be concluded that the root of the decision tree would be the same parameter. In Figure 2, as can be observed, the results of the decision tree classification are not completely consistent with logical thinking, and in some cases, they seem contradictory.

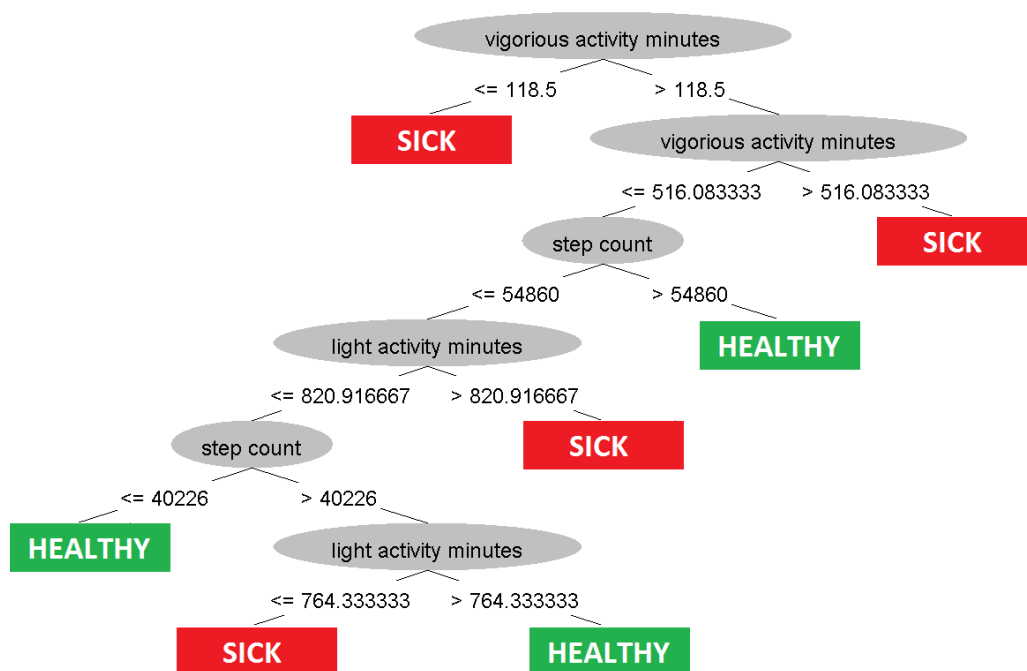


Figure 2. J48 tree design.

Accurate and reliable information is vital for effective decision making. Thus, we employed an undersampling technique to obtain reliable estimates. It is a technique used to adjust the class distribution of a dataset. For this purpose, 115 of the 215 sick children were chosen by the random selection process to obtain two equivalent ratios of sick and healthy patient classes. After undersampling, the remaining 230 patients were considered eligible and were enrolled in the study.

The best results in the prediction of type 1 diabetes presence among children and adolescents were obtained with decision tree forests. This model enabled the prediction with the highest accuracy (86.09%), specificity (84.35%), and precision (84.87%). The PNN also showed high accuracy (84.35%) and the highest sensitivity (89.57%), but markedly lower specificity (79.13%) and precision (81.10%). The AUC for PNN (0.926578) also exceeded the values of this parameter for the remaining classifiers (Figure 1). The averaged accuracy, sensitivity, specificity, precision, goodness index, and AUC value obtained for all applied computational intelligence methods and a linear regression model are presented in Table 4.

**Table 4.** The accuracy, sensitivity, specificity, precision, goodness index, and the area under the receiver operating characteristic curve obtained for the set of physical activity variables.

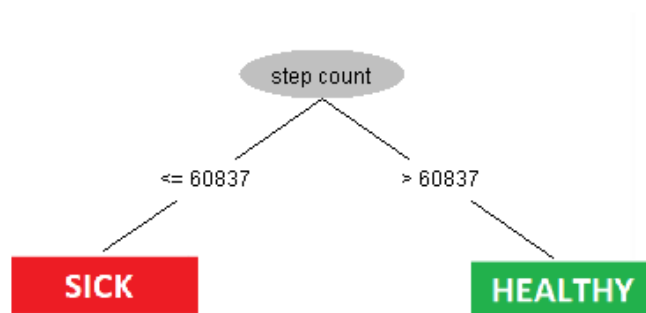
Algorithm Name	Acc(%)	Sen(%)	Spe(%)	Prec(%)	G	AUC
<b>Decision Tree Forest</b>	86.09	87.83	84.35	84.87	0.1983	-
PNN	84.35	89.57	79.13	81.10	0.2333	0.926578
SVM	84.35	86.96	81.74	82.64	0.2244	0.909716
Single tree	83.48	86.09	80.87	81.82	0.2365	-
GEP	83.04	83.48	82.61	82.76	0.2399	0.830435
Logistic regression	82.61	84.35	80.87	81.51	0.2472	0.883478
GMDH	82.61	82.61	82.61	82.61	0.2460	0.905482
RBF network	82.17	85.22	79.13	80.33	0.2557	0.905331
MLP	81.30	86.09	76.52	78.57	0.2729	0.897921
Linear regression	80.87	85.22	76.52	78.40	0.2774	0.884008

The values were obtained using a 10-fold cross-validation procedure [49]. The given dataset consisting of 230 samples was split into 10 folds, where each fold was used as a testing set at some point. In the first iteration, the first fold was used to test the model, and the rest were used to train the model. In the second iteration, the second fold was used as the testing set, and the rest served as the training set. This process was repeated until each fold of the 10 folds had been used as the testing set. Based on the obtained scores in every iteration, the mean value was calculated in order to assess the performance of the model.

### 4.3. Clustering Result

In the last step, we wish to explain the correlation between physical activity parameter values and type 1 diabetes presence. This relied on finding the equation that played a major role in the correct diabetes classification among children and adolescents. For this purpose, we assumed that we did not have a classification into sick and healthy patients and used the k-means clustering algorithm.

After clustering, we compared the obtained clusters with their corresponding classes from the dataset. It turned out that 215 of 330 records had identical classes. Then, we built a decision tree for the remaining 215 records using the c4.5 algorithm with the same setup as described in the Classification Result section. The result of the decision tree is presented in Figure 3.



**Figure 3.** Classification results after clustering.

The obtained results confirmed the assumption that the correlation between physical activity and type 1 diabetes presence can be evaluated based on measuring step count. It is possible to predict the prevalence of the disease correctly at least in 65% of the cases. As a result, a child was determined to be sick when performing fewer than 60,837 steps per week.

## 5. Discussion

The purpose of this research was to find a relationship between the intensity of physical activity and the presence of type 1 diabetes among children and develop a non-invasive method of type 1 diabetes detection. Assessment of the physical activity was based on ActiGraph activity monitor measurements. The ActiGraph measurements for health-related research were also carried out in published findings [17–20].

Decision tree forests, as well as other computational intelligence methods were applied for the detection of different diseases, e.g., breast cancer and heart disease [50,51]. Application of decision tree forests, which included five parameters connected with the intensity of physical activity, enabled the prediction of type 1 diabetes presence among children and adolescents between the ages of six and 18 with a high accuracy of 86.09% and specificity of 84.35%. The PNN also showed a high accuracy of 84.35%. Our results were comparable to similar articles, in which neural networks were used for outcome prediction of diabetes presence. For example, an SVM algorithm using the RBF kernel, the same as was used in this paper, was able to predict the presence of elevated blood glucose level via electrochemical measurement of saliva with approximately 85% accuracy [52].

As the final result of the study, it was concluded that if the number of steps is lower than around 61,000 a week, it is likely that the child is suffering from type 1 diabetes. After dividing this by the seven days of a week, we obtained the average number of steps per day, which was around 9000, but it should be noted that gender and age were not included in the calculation of this result. The updated international literature indicates that we can expect, among children, boys to average 12,000–16,000 steps/day, girls to average 10,000–13,000 steps per day, and adolescents to reach approximately 8000–9000 steps/day [53]. Thus, the obtained result was consistent with the normative international literature.

Decreased physical activity of ill children compared to healthy peers was the result of the disease. Many children also complained that it was difficult for them to go through a pitch of more than 100 m, that it was difficult for them to run, play sports, exercise, lift something heavy, take a bath, or shower by themselves, and that they felt pain and were tired.

The results of the research are promising and encourage developing a mobile application for type 1 diabetes diagnosis dedicated to children and adolescents. Although the popularity of using mobile phones applications in various health disorders has reached about 30%, it should be taken into account that young people are more likely to use and more effective at using new mobile phone applications, and the popularity and potential acceptance of mobile health solutions have an increasing tendency [54].

**Author Contributions:** Conceptualization, A.C., S.C., and D.M.; data curation, S.C.; formal analysis, S.C.; funding acquisition, D.M.; investigation, S.C.; methodology, A.C. and S.C.; project administration, A.C.; software, S.C.; supervision, D.M.; validation, D.M.; visualization, S.C.; writing, original draft, A.C.; writing, review and editing, A.C., S.C., and D.M.

**Funding:** This project is financed by the Minister of Science and Higher Education of the Republic of Poland within the “Regional Initiative of Excellence” program for years 2019–2022; Project Number 027/RID/2018/19; the amount granted: 11,999,900 PLN.

**Acknowledgments:** There are no acknowledgments related to this study.

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

## Abbreviations

The following abbreviations are used in this manuscript:

ADA	American Diabetes Association
AUC	Area under the receiver operating characteristic curve
BMI	Body mass index
DT	Decision tree
EE	energy expenditure
FN	False negative
FP	False positive
FR	Feature ranking
G	Goodness index
GEP	Gene expression programming
GMDH	Group method of data handling
LPA	light physical activity
MLP	Multilayer perceptron
MPA	moderate physical activity
PNN	Probabilistic neural network
RBF	Radial basis function
RF	Random forest
SVM	Support vector machine
TN	True negative
TP	Truth positive
WHO	World Health Organization
VPA	vigorous physical activity

## References

1. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **2009**, *33* (Suppl. 1), 62–67.
2. Tatoń, J.; Czech, A.; Bernas, M. Edukacja terapeutyczna, samokontrola glikemii i psychologia cukrzycy. Terapeutyczny styl życia. In *Diabetologia Kliniczna*; Tatoń, J., Czech, A., Bernas, M., Eds.; PZWL: Warsaw, Poland, 2008; pp. 339–429.
3. World Health Organization. *Global Strategy on Diet, Physical Activity and Health*; WHO Library Cataloguing-in-Publication Data; World Health Organization: Geneva, Switzerland, 2004; pp. 1–18.
4. Iannotti, R.J.; Kalman, M.; Inchley, J.; Tynjälä, J.; Bucksch, J.; The HBSC Physical Activity Focus Group. Social determinants of health and well-being among young people. In *Health Behaviour in School-Aged Children (HBSC) Study: International Report from the 2009/2010 Survey*; Currie, C., Ed.; WHO Regional Office for Europe: Copenhagen, Denmark, 2012; pp. 129–132.
5. Faigenbaum, A. Physical Activity in Children and Adolescents. *ACSM Bull.* **2015**. Available online: <https://www.acsm.org/> (accessed on 16 October 2018).
6. International Diabetes Federation. *IDF Diabetes Atlas. Eighth Edition 2017*; International Diabetes Federation: Brussels, Belgium, 2017; Volume 8, pp. 1–150.
7. Pettitt, D.J.; Talton, J.; Dabelea, D.; Divers, J.; Imperatore, G.; Lawrence, J.M.; Liese, A.D.; Linder, B.; Mayer-Davis, E.J.; Pihoker, C.; et al. Prevalence of Diabetes in U.S. Youth in 2009: The SEARCH for Diabetes in Youth Study. *Diabetes Care* **2014**, *37*, 402–408. [[CrossRef](#)] [[PubMed](#)]
8. Czenczek-Lewandowska, E. Level of Physical Activity in Children and Adolescents with type 1 Diabetes, Relative to the Insulin Therapy Applied. Ph.D. Thesis, University of Rzeszów, Rzeszów, Poland, 2017; pp. 1–165.
9. Czenczek-Lewandowska, E.; Grzegorzcyk, J.; Mazur, A. Physical activity in children and adolescents with type 1 diabetes and contemporary methods of its assessment. *Pediatr. Endocrinol. Diabetes Metab.* **2018**, *24*, 179–184. [[CrossRef](#)] [[PubMed](#)]
10. Allen, N.; Gupta, A. Current Diabetes Technology: Striving for the Artificial Pancreas. *Diagnostics* **2019**, *9*, 31. [[CrossRef](#)] [[PubMed](#)]
11. Strath, S.J.; Kaminsky, L.A.; Ainsworth, B.E.; Ekelund, U.; Freedson, P.S.; Gary, R.A.; Richardson, C.R.; Smith, D.T.; Swartz, A.M. Guide to the Assessment of Physical Activity: Clinical and Research Applications A Scientific Statement From the American Heart Association. *Circulation* **2013**, *128*, 2259–2279. [[CrossRef](#)]

12. Katch, V.L.; McArdle, W.D.; Katch, F.I. Energy expenditure during rest and physical activity. In *Essentials of Exercise Physiology*, 4th ed.; McArdle, W.D., Katch, F.I., Katch, V.L., Eds.; Lippincott Williams & Wilkins: Baltimore, MD, USA, 2011; pp. 237–262.
13. Sylvia, L.G.; Bernstein, E.E.; Hubbard, J.L. Practical Guide to Measuring Physical Activity. *J. Acad. Nutr. Diet.* **2014**, *114*, 199–208. [[CrossRef](#)]
14. Hills, A.P.; Mokhtar, N.; Byrne, N.M. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. *Front. Nutr.* **2014**, *1*, 1–14. [[CrossRef](#)]
15. Tanaka, C.; Hikihara, Y.; Ando, T.; Oshima, Y.; Usui, C.; Ohgi, Y.; Kaneda, K.; Tanaka, S. Prediction of Physical Activity Intensity with Accelerometry in Young Children. *Int. J. Environ. Res. Public Health* **2019**, *16*, 931. [[CrossRef](#)]
16. van Hees, V.T.; Pias, M.; Taherian, S.; Ekelund, U.; Brage, S. A method to compare new and traditional accelerometry data in physical activity monitoring. In Proceedings of the 2010 IEEE International Symposium on “A World of Wireless, Mobile and Multimedia Networks”, Montreal, QC, Canada, 14–17 June 2010; pp. 1–6.
17. Vijay, R.; Watts, A.; Watts, V. Daily Physical Activity Patterns During the Early Stage of Alzheimer’s Disease. *J. Alzheimer’s Dis.* **2016**, *55*, 659–667.
18. Bonato, P.; Sherrill, D.M.; Standaert, D.G.; Salles, S.S.; Akay, M. Data mining techniques to detect motor fluctuations in Parkinson’s disease. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2004**, *7*, 4766–4769. [[PubMed](#)]
19. Ahmadi, M.; O’Neil, M.; Fragala-Pinkham, M.; Lennon, N.; Trost, S. Machine learning algorithms for activity recognition in ambulant children and adolescents with cerebral palsy. *J. Neuroeng. Rehabil.* **2018**, *15*, 105. [[CrossRef](#)] [[PubMed](#)]
20. Quante, M.; Cespedes Feliciano, E.M.; Rifas-Shiman, S.L.; Mariani, S.; Kaplan, E.R.; Rueschman, M.; Oken, E.; Taveras, E.M.; Redline, S. Association of Daily Rest-Activity Patterns With Adiposity and Cardiometabolic Risk Measures in Teens. *J. Adolesc. Health* **2019**. [[CrossRef](#)] [[PubMed](#)]
21. Kanna, K.R.; Sugumaran, V.; Vijayaram, T.R.; Karthikeyan, C.P. Activities of Daily Life (ADL) Recognition using Wrist-worn Accelerometer. *Int. J. Eng. Technol. (IJET)* **2016**, *8*, 1406–1413.
22. Welk, G.J. Use of accelerometry-based activity monitors to assess physical activity. In *Physical Activity Assessments for Health-Related Research*; Welk, G.J., Ed.; Human Kinetics Publishers: Champaign, IL, USA, 2002; pp. 125–142.
23. Crouter, S.E.; Horton, M.; Bassett, D.R. Validity of ActiGraph Child-Specific Equations during Various Physical Activities. *Med. Sci. Sports Exerc.* **2013**, *45*, 1403–1409. [[CrossRef](#)] [[PubMed](#)]
24. Hekler, E.B.; Buman, M.P.; Grieco, L.; Rosenberger, M.; Winter, S.J.; Haskell, W.; King, A.C. Validation of Physical Activity Tracking via Android Smartphones Compared to ActiGraph Accelerometer: Laboratory-Based and Free-Living Validation Studies. *JMIR MHealth UHealth* **2015**, *3*, e36. [[CrossRef](#)] [[PubMed](#)]
25. Migueles, J.H.; Cadenas-Sanchez, C.; Ekelund, U.; Delisle Nyström, C.; Mora-Gonzalez, J.; Löf, M.; Labayen, I.; Ruiz, J.R.; Ortega, F.B. Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Med.* **2017**, *47*, 1821–1845. [[CrossRef](#)] [[PubMed](#)]
26. Jacob, E. Classification and Categorization: A Difference that Makes a Difference. *Libr. Trends* **2004**, *52*, 515–540.
27. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of Support Vector Machine(SVM) Learning in Cancer Genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51.
28. Taborri, J.; Palermo, E.; Rossi, S. Automatic Detection of Faults in Race Walking: A Comparative Analysis of Machine-Learning Algorithms Fed with Inertial Sensor Data. *Sensors* **2019**, *19*, 1461. [[CrossRef](#)]
29. Sun, Q.; Lin, F.; Yan, W.; Wang, F.; Chen, S.; Zhong, L. Estimation of the Hydrophobicity of a Composite Insulator Based on an Improved Probabilistic Neural Network. *Energies* **2018**, *11*, 2459. [[CrossRef](#)]
30. Nazzal, J.M.; El-Emary, I.M.; Najim, S.A. Multilayer Perceptron Neural Network (MLPs) For Analyzing the Properties of Jordan Oil Shale. *World Appl. Sci. J.* **2008**, *5*, 546–552.
31. Li, R.Y.M.; Fong, S.; Chong, W.S. Forecasting the REITs and stock indices: Group Method of Data Handling Neural Network approach. *Pac. Rim Prop. Res. J.* **2017**, *23*, 1–38. [[CrossRef](#)]
32. Ferreira, C. The Basic Gene Expression Algorithm. In *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*; Springer: Berlin, Germany, 2006; pp. 55–120.

33. Godfrey, K. Simple Linear Regression in Medical Research. *N. Engl. J. Med.* **1985**, *313*, 1629–1636. [[CrossRef](#)] [[PubMed](#)]
34. Acosta, F.M.A. Radial basis function and related models: An overview. *Signal Process.* **1995**, *45*, 37–58. [[CrossRef](#)]
35. Peng, C.Y.J.; Lee, K.L.; Ingersoll, G.M. An Introduction to Logistic Regression Analysis and Reporting. *J. Educ. Res.* **2002**, *96*, 3–14. [[CrossRef](#)]
36. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Tree-Based Methods. In *An Introduction to Statistical Learning with Applications in R*; Springer: New York, NY, USA, 2017; pp. 303–336.
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
38. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [[CrossRef](#)]
39. Brisimi, T.; Xu, T.; Wang, T.; Dai, W.; Adams, W.; Paschalidis, I. Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach. *Proc. IEEE* **2018**, *106*, 690–707. [[CrossRef](#)]
40. Taborri, J.; Scalona, E.; Palermo, E.; Rossi, S.; Cappa, P. Validation of Inter-Subject Training for Hidden Markov Models Applied to Gait Phase Detection in Children with Cerebral Palsy. *Sensors* **2015**, *15*, 24514–24529. [[CrossRef](#)]
41. Wilkin, G.A.; Huang, X. K-Means Clustering Algorithms: Implementation and Comparison. In *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*; IEEE: Iowa City, IA, USA, 2007.
42. Cilia, N.; De Stefano, C.; Fontanella, F.; Raimondo, S.; di Freca, A.S. An Experimental Comparison of Feature-Selection and Classification Methods for Microarray Datasets. *Information* **2019**, *10*, 109. [[CrossRef](#)]
43. Rodgers, J.L.; Nicewander, W.A. Thirteen Ways to Look at the Correlation Coefficient. *Am. Stat.* **1988**, *42*, 59–66. [[CrossRef](#)]
44. Robert, C. An entropy concentration theorem: Applications in artificial intelligence and descriptive statistics. *J. Appl. Probab.* **1990**, *27*, 303–313. [[CrossRef](#)]
45. Quinlan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [[CrossRef](#)]
46. Sherrod, P. DTREG Predictive Modeling Software. 2003. Available online: [www.dtreg.com](http://www.dtreg.com) (accessed on 12 February 2019).
47. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [[CrossRef](#)]
48. Michel, V.; Gramfort, A.; Varoquaux, G.; Eger, E.; Keribin, C.; Thirion, B. A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognit.* **2012**, *45*, 2041–2049. [[CrossRef](#)]
49. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. Int. Jt. Conf. Artif. Intell.* **1995**, *14*, 1137–1143.
50. Übeyli, E.D. Implementing automated diagnostic systems for breast cancer detection. *Expert Syst. Appl.* **2007**, *33*, 1054–1062. [[CrossRef](#)]
51. Nahar, J.; Imam, T.; Tickle, K.S.; Chen, Y.P. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Syst. Appl.* **2013**, *40*, 96–104. [[CrossRef](#)]
52. Malik, S.; Khadgawat, R.; Anand, S.; Gupta, S. Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. *SpringerPlus* **2016**, *5*, 701. [[CrossRef](#)]
53. Tudor-Locke, C.; Craig, C.L.; Beets M.W.; Belton, S.; Cardon, G.M.; Duncan, S.; Hatano, Y.; Lubans, D.R.; Olds, T.S.; Raustorp, A.; et al. How many steps/day are enough? for children and adolescents. *Int. J. Behav. Nutr. Phys. Act.* **2011**, *8*, 78. [[CrossRef](#)] [[PubMed](#)]
54. Cerna, L.; Maresova, P. Patients' attitudes to the use of modern technologies in the treatment of diabetes. *Patient Prefer Adherence* **2016**, *10*, 1869–1879. [[PubMed](#)]

