electronics

MDPI

Article

RG Hyperparameter Optimization Approach for Improved Indirect Prediction of Blood Glucose Levels by Boosting Ensemble Learning

Yufei Wang [1], Haiyang Zhang [2], Yongli An [1], Zhanlin Ji [1,3,*] and Ivan Ganchev [3,4,5,*]

[1] College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China; yufi_w@163.com (Y.W.); anyl@ncst.edu.cn (Y.A.)
[2] Department of Computer Science, University of Sheffield, Sheffield S10 2TN, UK; haiyang.zhang@sheffield.ac.uk
[3] Telecommunications Research Centre (TRC), University of Limerick, V94 T9PX Limerick, Ireland
[4] Department of Computer Systems, University of Plovdiv "Paisii Hilendarski", 4000 Plovdiv, Bulgaria
[5] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria
* Correspondence: zhanlin.ji@ncst.edu.cn (Z.J.); ivan.ganchev@ul.ie (I.G.)

Abstract: This paper proposes an RG hyperparameter optimization approach, based on a sequential use of random search (R) and grid search (G), for improving the blood glucose level prediction of boosting ensemble learning models. An indirect prediction of blood glucose levels in patients is performed, based on historical medical data collected by means of physical examination methods, using 40 human body's health indicators. The conducted experiments with real clinical data proved that the proposed RG double optimization approach helps improve the prediction performance of four state-of-the-art boosting ensemble learning models enriched by it, achieving 1.47% to 24.40% MSE improvement and 0.75% to 11.54% RMSE improvement.

Keywords: blood glucose level; prediction; ensemble learning; boosting; hyperparameter optimization; random search; grid search

check for updates

Citation: Wang, Y.; Zhang, H.; An, Y.; Ji, Z.; Ganchev, I. RG Hyperparameter Optimization Approach for Improved Indirect Prediction of Blood Glucose Levels by Boosting Ensemble Learning. *Electronics* **2021**, *10*, 1797. https://doi.org/10.3390/electronics10151797

Academic Editors: Juan Manuel Corchado Rodriguez, Chilukuri Krishna Mohan, Luca Mesin, Christian Morbidoni, Francesco Di Nardo and Alessandro Cucchiarelli

Received: 3 April 2021
Accepted: 23 July 2021
Published: 27 July 2021

1. Introduction

Diabetes mellitus is a chronic non-communicable disease, which is closely related to people's dietary habits and lifestyle. As of 2019, the estimated number of people with diabetes has reached 463 million and there were an estimated 4.2 million deaths among adults (aging from 20 to 79 years) attributable to diabetes worldwide [1], and these numbers continue to grow. The authors in [1] have also found that "Excess glucose has been shown to be associated with about 15% of all deaths due to CVD (CVD stands for cardiovascular disease), kidney disease, and diabetes . . . , indicating a large number of these premature deaths can be potentially prevented through prevention or early detection of type 2 diabetes mellitus (the three common types of diabetes mellitus include: *type 1*, when the human body fails to produce insulin; *type 2*, when the cells fail to use insulin; and *gestational*, with high blood sugar level during pregnancy [2]) and improved management of all forms of diabetes and these complications." As an important health problem existing in many countries, e.g., China [3], diabetes requires continuous surveillance and effective control for tackling it properly.

One way to achieve this is to make full use of the medical history of people, e.g., obtained through regularly performed comprehensive physical examinations, established as a routine practice, especially in the developed countries. In China, for instance, there are many hospitals where such examinations can be done without a prior appointment, with the cost of the examination being less than one tenth of the average monthly income. The only requirement imposed for people is to appear in the hospital on an empty stomach in the morning. In addition, each hospital has a special physical examination center, where companies arrange regularly one or two cost-free physical examination(s) of their employees every year to master their health status. A typical medical examination includes checking the liver functioning, blood fat, kidney functioning, hepatitis B virus existence, blood routine examination, electrocardiograph, chest X-raying, B-mode ultrasound image, etc. However, in the physical examination program in China, the blood routine examination is mainly focused on blood cytology and cell morphology, and the typical instrument used is a blood cell analyzer [4]. To check the blood glucose level, nurses need to take blood again, which brings extra pain to physical examinees. Finally, the disposal of medical waste is also an issue.

Thus, people seem to pay more attention to the use of non-invasive methods for the prediction of blood glucose levels, but currently only optical technology seems to have a good development prospect [5,6]. Others, such as thermal-, electrical-, and nanotechnology methods, are still theoretical. However, optical techniques still have many limitations in predicting blood glucose levels. For example, the intermediates used in fluorescence technology are toxic [7], which may harm the person being tested, and, in addition, the sensors have a short service life. The disadvantages of mid-infrared spectroscopy (MIRS) [6], such as poor penetration and expensive equipment, must also be taken into account. Other methods such as optical polarimetry [8] and optical coherence tomography [5] are very sensitive to temperature. Wearable dynamic blood glucose monitors that use body fluids [9] may be a good alternative, but they are not yet on the market in large numbers and their cost is not affordable for every family. Therefore, non-invasive methods for the prediction of blood glucose levels have not been widely used.

Research scholars nowadays focus strongly on the use of the full medical history of people, e.g., obtained through regular physical examinations, to predict their blood glucose levels. First of all, this is due to the fact that point-of-care glucose meters use different measurement methods leading to device-specific limitations, interferences, and technical constraints [10]. Secondly, the device type, sampling conditions, and interpretation of results must also be taken into consideration. For example, in order to facilitate patients to keep eye on their blood sugar level, some self-testing devices bring convenience to patients without the help of professionals. However, all the testing equipment on the market needs a test paper, which, after reacting with the oxygen in the air, may yield incorrect results. In addition, the test paper and the blood glucose meter must be produced by the same manufacturer, which brings unnecessary trouble and less freedom of choice to patients.

To improve the prediction of blood glucose levels in patients, based on their historical medical data, this paper follows the idea, presented in [11], of using multiple human body's health indicators, collected by means of regular physical examinations and processed by machine learning (ML) techniques. However, instead of the HY_LightGBM model proposed in [11], other state-of-the-art ML models, namely boosting ensemble learning models, are utilized here, all enriched by the proposed RG hyperparameter double optimization by means of random search (R) and grid search (G). The results, obtained from the conducted experiments, confirmed that the proposed RG double optimization helps improve the blood glucose level prediction of the considered state-of-the-art boosting ensemble learning models, enriched by this RG approach, while also outperforming the HY_LightGBM model proposed in [11].

The research, reported here, explores the relationship between the blood glucose and other human body's health indicators. The presented study uses biochemical data of liver functioning, kidney functioning, blood routine, etc., to explore the relationship between the blood glucose and such data, showing that this could be used for indirect prediction of blood glucose levels. Numerous results reported in the literature confirm that the biochemical data, utilized in this research, are indeed related to the blood glucose, and thus one can infer blood glucose levels from such data. For instance, the authors in [12] show that the odds ratio of developing type 2 diabetes rises significantly with increasing the levels of serum liver enzymes, γ-glutamyl transferase (GGT) and alanine aminotransferase (ALT), i.e., two of the 40 human body's health indicators utilized in the study reported here. The same authors conclude that increased GGT and ALT levels are independent, additive risk factors for the development of type 2 diabetes mellitus in subjects without fatty liver or hepatic dysfunction. In [13], the GGT and ALT levels were found to be closely related to prediabetes and diabetes in overweight and obese people, and positively associated with insulin resistance. In [14], the GGT level was reported as a significant predictor of subsequent risk of diabetes mellitus, increased by 4% for every 1 IU/L increase in GGT (<24 IU/L). A study on the relation of liver enzymes with the development of type 2 diabetes, presented in [15], suggests that ALT concentrations are independently associated with type 2 diabetes in both males and females, whereas the GGT level is also independently associated but only for females (sex of patients was also taken into account by the research presented here). In [16], the liver enzymes were also found independent risk factors for elevated blood glucose, with presented sex differences in the role of each enzyme. The research results reported in [17] show that, among others, age and serum triglyceride (TG)—i.e., another two human body's health indicators considered by the research presented here—are directly related to risk of type 2 diabetes. Moreover, the authors of [17] saw similar gradients for diabetes across fitness groups in strata of high-density lipoprotein cholesterol level (TC), which is another human body's health indicator utilized in the study presented here. The authors in [18] pointed out that, among other factors, increased concentration of low-density-lipoprotein cholesterol (LDL_C) and decreased concentration of high-density lipoprotein cholesterol (HDL_C)—another two human body's health indicators included in the study presented here—are the strongest risk factors for patients with type 2 diabetes. In addition, these authors underlined that high concentrations of triglyceride—yet another human body's health indicator utilized in the research presented here—are typically observed in people with type 2 diabetes. In [19], the increased ratio of triglyceride to HDL_C has been associated with an increased risk of all-cause and cardiovascular mortality in type 2 diabetic subjects, largely mediated by the presence of kidney dysfunction. As stated in [20], the inverse relationship between LDL_C and diabetes has been confirmed by multiple clinical trials and genetic instruments using aggregate single nucleotide polymorphisms. In addition, at least eight individual genes support this inverse association. Moreover, genetic and pharmacologic evidence suggest that HDL_C may also be inversely associated with risk for diabetes. As stated in [21], HDL_C, triglyceride, and total cholesterol (TC)—used as human body's health indicators in the presented here research– are identified as the top three most consistent predictors of a coronary heart disease in type 2 diabetes subjects. Further on, the authors of [21] found a significant positive linear correlation between elevated blood glucose and total cholesterol, triglycerides, and LDL. The same authors conclude that type 2 diabetes mellitus is strongly associated with lower level of HDL_C and higher level of LDL_C.

It should be specially noted that the goal of the research reported in this paper was not to replace the routine blood glucose testing program, carried out in hospitals, with machine learning techniques, but rather to explore the relationship between the blood glucose level and other health indicators of human body that are obtained by periodic tests which, however, do not include blood glucose level's examination. In such a context, the proposed approach can provide an early alert so that unsuspected diabetic cases can be identified as early as possible in order to start treating them promptly. Such research belongs to the interdisciplinary field of medical research and data science, as revealed above.

The rest of the paper is organized as follows. The next section presents the related work done in this field, whereas Section 3 describes the background. Section 4 explains the proposed RG hyperparameter optimization approach. Section 5 presents the experimental performance evaluation of the compared models and discusses the results. Finally, Section 6 concludes the paper and sets future directions for research.

2. Related Work

Machine learning (ML) has achieved very good results for prediction and timely treatment of various diseases [22,23]. For instance, Solanki et al. [24], proposed methods for improving the performance of ML classification models, namely a support vector machine (SVM), a decision tree, and a multilayer perceptron (MLP), i.e., a feed-forward artificial neural network (ANN), for the prognosis of breast cancer. ML models can be utilized also to predict blood glucose levels, based on collected medical data and various human body's health indicators. The MLP diabetes prediction expert system, designed by Jahangir et al. [3], performed an outlier detection of data before making a prediction, with accuracy of 88.7%. Santhanam et al. [25] used a K-means clustering algorithm to remove the noise in Pima-Indians data, found the characteristic value by a genetic algorithm, and finally brought it into a SVM classifier to determine whether the test population had diabetes. However, this study did not process the missing data and outliers, which would otherwise have allowed it to increase the accuracy. Nai-arun et al. [26] analyzed a real data set, collected from a hospital in Thailand, using the integration idea and performed bagging and boosting fusion separately using Naïve Bayes, K nearest neighbors (KNN), and decision trees as base classifiers. The bagging approach demonstrated an accuracy of 95.3% for the base classifier fusion, which indicated that the use of the integration approach had a better predictive effect than applying the model alone. Wang et al. [11] proposed the HY_LightGBM model, utilizing a Bayesian optimization algorithm for finding the optimal values of hyperparameters, for predicting the blood glucose levels, showing that their model outperforms the XGBoost model [27], the LightGBM model [28] optimized by a genetic algorithm, and the LightGBM model optimized by a random search.

Following the idea presented in [11] of using clinical data and human body's health indicators, obtained by physical examination of patients in a tertiary-care hospital, for predicting their blood glucose levels, an RG hyperparameter optimization approach is proposed in this paper for improving the prediction of boosting ensemble learning models. However, as some clinical data in the utilized data set were missing, the importance of features with missing data is first analyzed in order to conclude whether some of these have value, which are subsequently not deleted, as done in [11], but filled with the medians. In addition, in order to avoid poor prediction on normal data, the outlier data are first removed using boxplots and then substituted with the medians. A final strong learner is generated by using a residual iteration and fitting a regression tree.

Grid search is a commonly used hyperparameter adjustment method. Its principle is to combine all possible hyperparameters and cycle each hyperparameter combination until the best combination is found. Although this method is simple and easy to implement, its use may cause waste of computing resources and time, especially in models working with many hyperparameters, such as GBDT [29]. Aiming at the shortcomings of the grid search, Bergstra et al. [30] proposed the use of random search to find the hyperparameters' values by randomly sampling the hyperparameters in a limited range. The hyperparameters of continuous variables are regarded as a distribution for sampling. Therefore, the method can quickly determine the approximate range of hyperparameters. By sequentially applying these two methods, the RG optimization approach, proposed in this paper, first avails of a random search (R) for determining the approximate range of the hyperparameters, followed by a grid search (G) for finding their optimal values within this range.

In the past, scholars have predicted diabetes using ANNs, or a single learner, with poor interpretability or unsatisfactory prediction results, while ensemble learning models based on boosting (e.g., AdaBoost [31], GBDT, XGBoost, LightGBM) allow to greatly reduce the prediction error through continuous fitting of residual errors. The prediction error of these models could be further reduced by a sequential use of random search and grid search, as demonstrated further in this paper. Thanks to this RG double optimization applied to the state-of-the-art boosting ensemble learning models, their prediction performance can be improved (quite significantly in some cases). Thus, the proposed RG optimization approach can help better predict the patient's blood glucose levels, avoid errors caused by human factors, improve the work efficiency of the healthcare providers, and compensate the deficiency of the existing boosting ensemble learning models used for prediction of diabetes.

3. Background

3.1. Ensemble Learning Models

Ensemble learning is a powerful ML paradigm whereby multiple learners are trained for solving the same problem, such as text categorization, optical character recognition, face recognition, gene expression analysis, computer-aided medical diagnosis, etc. [32]. Instead of trying to learn one hypothesis from the training data, as in the ordinary ML, ensemble learning tries to construct a set of hypotheses for combined use. An ensemble contains a few learners, called base learners or weak learners, which are generated from the training data by means of a single base learning algorithm (e.g., a decision tree, an ANN, etc.) or multiple algorithms. Then, the base learners are combined for use, e.g., by means of weighted averaging in the case of solving a regression problem, or majority voting in the case of a classification problem. The use of multiple learners helps ensemble learning get much better generalization ability than that of a single learner. After proving made by Schapire in 1989 [33] that weak learners can be boosted to strong learners, *boosting* has emerged as one of the most influential ensemble learning approaches (the other two are *bagging* and *stacking*). Boosting often does not suffer from overfitting even after a large number of rounds, and sometimes it is even able to reduce the generalization error after the training error reaching zero. Moreover, in addition to reducing the variance, boosting can significantly reduce the bias, and thus, on weak learners, it is usually more effective [32]. The main representatives of boosting ensemble learning models are briefly described in the following subsections.

3.1.1. AdaBoost

The adaptive boosting (AdaBoost) model was developed by Freund and Schapire [31] in 1997. After initially assigning equal weights to all training examples, it generates a base learner from the training data set by calling the base learning algorithm [32]. Then, it uses the training examples to test the base learner and increases the weights of the incorrectly classified examples. From the training data set and updated weight distribution, AdaBoost generates another base learner by calling the base learning algorithm again. After repeating this process R rounds, AdaBoost derives the final learner by weighted majority voting of the R base learners. In practice, the base learning algorithm may use weighted training examples directly, or otherwise the weights can be exploited by sampling the training examples according to the weight distribution [32].

3.1.2. GBDT

Gradient boosting decision tree (GBDT), otherwise known as multiple additive regression tree (MART), is an iterative decision tree based model. It differs from AdaBoost, which adjusts the weight according to the classification effect and then iterates continuously. Instead, GBDT iterates with the negative gradient of the loss function as the approximation of the residual, fits the regression tree, and finally forms a strong learner. The idea is to combine multiple decision trees together to produce the final result. Although GBDT can also be classified, its decision tree is a regression tree, so its core lies in accumulation, that is, summing up the conclusions of all trees to reach the final conclusion. In other words, the input of each tree learning is the residual of the sum of all previous tree conclusions. The idea of gradient descent is used to calculate the residual. Freidman et al. [29] used the direction of negative gradient of the loss function to replace the direction of residual, so the negative gradient of the loss function is called pseudo residual. The direction of the pseudo residual is the locally optimal direction. The negative gradient of the loss function is used to fit the approximation of the current loss even if the loss function iterates to the minimum.

3.1.3. XGBoost

Developed by Chen et al. [27], extreme gradient boosting (XGBoost) is a model for a massively parallel boosted tree. The basic idea is the same as that of GBDT, which is based on the direction of the negative gradient of the loss function. However, the XGBoost's loss function is the second-order Taylor expansion of the error part. The regularization term is added to prevent overfitting, and the objective function of iterative optimization could be customized, if it is second-order differentiable. For large data sets, XGBoost consumes more memory and takes more execution time, as stated in [11], than the LightGBM model presented next.

3.1.4. LightGBM

LightGBM was proposed by Microsoft in 2017 [28]. Similarly to XGBoost, it supports parallel arithmetic, but is more powerful and can be trained faster [11]. It is featured by a decision tree algorithm based on: (i) a gradient-based one-side sampling (GOSS) for keeping all large gradient samples and performing random sampling on the small gradient samples; (ii) an exclusive feature bundling (EFB) for dividing the features into a smaller number of mutually exclusive bundles; and (iii) a histogram and leaf-wise growth strategy with a depth limit for finding a leaf node with the largest split gain in the current leaf nodes every time [11].

3.2. Other ML Models

3.2.1. HY_LightGBM

Although LightGBM can achieve high prediction performance, just like other boosting ensemble learning models, it involves many hyperparameters whose selection influences greatly the prediction results. Therefore, Wang et al. [11] have proposed a Bayesian hyperparameter optimization algorithm to determine the hyperparameter combination for use with LightGBM, which resulted in the HY_LightGBM model. In terms of data processing, the features with small missing values are filled with the medians, whereas the features with large missing values are simply deleted. This differs from the method used in this paper.

3.2.2. ANNs

Artificial neural networks (ANNs) abstract the human brain's neural network from the point of view of information processing. They are based on an interconnection of a large number of nodes, called artificial neurons, which are organized into multiple layers. The number of layers defines the depth of the network. Deep neural networks are generally used for image- and voice-processing, whereas shallow neural networks are more suitable for small-scale data sets.

ANNs are the most widely used classification and prediction ML tools at present, especially for data with high structure, e.g., voice, pictures, and natural languages [34]. However, the tree-based models have obvious advantages for small-scale data sets, because the increased complexity of the network can easily lead to overfitting in this case [35]. In addition, ANNs needs more rigorous data preparation, such as data type conversion, data standardization, etc. Moreover, the interpretation of an ANN model is far less convenient and less intuitive than the embedded feature selection of an integrated tree model. Therefore, the latter is usually superior to ANN when applied to small-scale data sets [35], such as those containing medical data.

To prove this, in the performance evaluation of models (c.f. Section 5), a specially designed and optimized ANN was included for comparison with the boosting ensemble learning models. This ANN consists of an input layer, three fully connected hidden layers, and an output layer (Figure 1). Experimentally, we found that the use of more than three hidden layers is not justified as it does not bring further improvement of accuracy and, in addition, it increases the running time and leads to overfitting. The ANN was trained and optimized by utilizing the adaptive learning rate algorithm ADAM [36]. Different from the ANN processing in case of image classification, the output layer of this ANN does not need to use an activation function and directly predicts the blood glucose level instead.

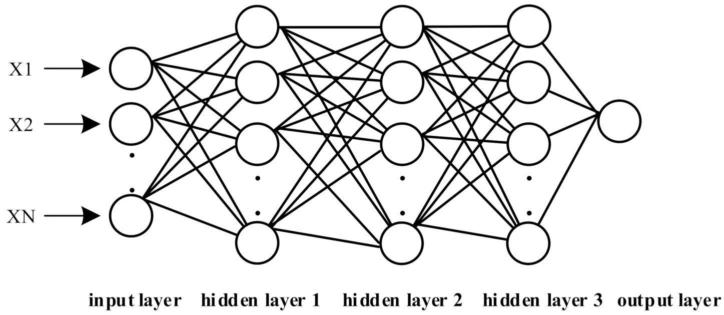

Figure 1. The ANN used in the experiments.

3.3. Loss Functions

Different loss functions could be used for solving different problems. For regression problems, the most used loss functions are briefly presented in the following subsections.

3.3.1. MSE

The mean square error (MSE) is defined, e.g., in [37], as:

$$L(y, f(x)) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 \tag{1}$$

where y_i $(i = 1,\ 2,\ \ldots,\ N)$ denote the actual values and $f(x_i)$ denote their predicted values.

The corresponding negative gradient is:

$$-\frac{\partial L}{\partial f(x)} = y - f(x) \tag{2}$$

3.3.2. MAE

The mean absolute error (MAE) is defined, e.g., in [38], as:

$$L(y, f(x)) = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)| \tag{3}$$

The corresponding negative gradient error is:

$$sign(y_i - f(x_i)) \tag{4}$$

A big problem with MAE relates to its constantly large gradient, which could lead to missing minima at the end of training using gradient descent. In this regard, MSE is more precise as its gradient decreases as the loss gets close to its minima [39].

3.3.3. Huber Loss

Huber loss is a compromise between MSE and MAE. It is defined in [40] as:

$$L(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, |y - f(x)| \le \beta \\ \beta\left(|y - f(x)| - \frac{\beta}{2}\right), |y - f(x)| > \beta \end{cases} \tag{5}$$

where β is the hyperparameter of Huber loss.

The corresponding negative gradient error is:

$$r(y_i, f(x_i)) = \begin{cases} y_i - f(x_i), |y - f(x)| \le \beta \\ \beta sign(y_i - f(x_i)), |y - f(x)| > \beta \end{cases} \tag{6}$$

Huber loss curves around the minima which decreases the gradient. In addition, it is more robust to outliers than MSE. However, its main problem is that a training of the hyperparameter β is needed, which is an iterative process [41].

The approach, presented in this paper, uses the MSE loss function.

3.4. Search Methods

3.4.1. Grid Search

Grid search [41] is a commonly used search method, but with a low-search efficiency. It requires the determination of L candidate values for each hyperparameter and a random combination of the candidate values of K hyperparameters to form alternative parameters. The number of experiments in grid search is:

$$S = \prod_{K=1}^{K} \left| L^{(K)} \right| \tag{7}$$

In this method, the growth of the number of hyperparameters may lead to a dimensional catastrophe and could also make the selection of these difficult. Moreover, for a large number of hyperparameters, grid search is very slow.

3.4.2. Random Search

Random search [30] uses a random number to obtain the optimal solution. This method continuously generates random points in a certain interval and calculates the values of a constraint function and an objective function. For the points meeting the constraint conditions, the values of the objective function are compared one by one, and the optimal values are saved. Instead of trying all possible combinations, the random search performs sampling according to the distribution of each hyperparameter and selects a specific number of hyperparameters for random combinations. However, the random search exhibits a poor performance when applied to small-scale data sets [26].

4. RG Hyperparameter Optimization Approach

The proposed RG hyperparameter optimization approach is based on the sequential use of random search (“R” in the approach’s name) and grid search (“G” in the approach’s name). Ensemble learning models usually involve many hyperparameters, whose values’ selection has great impact on the prediction performance. A reasonable set of hyperparameters can reduce the prediction error. Manual tuning is a method of repeated experiments that consumes a lot of time. At the same time, since the grid search will try every hyperparameter combination, it will be extremely slow in finding the hyperparameters when the number of these is more than three. In many cases, hyperparameters are not equally important. Random selection of parameter combinations in the hyperparameter space is faster than the grid search, but because it does not ensure that the optimal hyperparameter combinations are given, it is necessary also to apply a grid search after the random search to adjust the range near each hyperparameter.

Therefore, in the proposed RG hyperparameter optimization approach, a random search is used first, followed by a grid search, to determine the optimal values of hyperparameters, as shown in Table 1 for the RG_GBDT model, used here as an example of the RG-enriched boosting ensemble learning models.

Table 1. The optimal values of hyperparameters determined for the RG_GBDT model.

Hyperparameter	Default Value	Optimal Value (*After Random Search*)	Optimal Value (*After Grid Search*)	Implication on
n_estimators	100	69	67	The number of boosting stages that will be performed
learning_rate	0.1	0.09	0.07	How much the contribution of each tree will shrink
subsample	1	0.7	0.8	The subsampling ratio
min_samples_split	2	280	280	The minimum number of samples required to split an internal node
min_samples_leaf	1	440	440	The smallest possible record tree for a leaf
max_features	none	6	6	The number of features to consider when looking for the best cut
max_depth	3	17	17	The limit of the number of nodes in the tree

After the optimal value of each hyperparameter is determined, the strongest expression [42] is obtained according to the following Algorithm 1:

Algorithm 1 Training Algorithm

```
Input:$\{x_1, x_2, \ldots,\ x_N\}$, where for the given $d$ = 40 data features $x_i = \left(x_{i1}, x_{i2}, \ldots, x_{id}\right)$ and $i = 1,\ 2, \ldots, N$.
The maximum number of iterations for the model is $T$ and the loss function is $L$.
Result: The predicted glucose level.
Begin:
1. Initialize the loss function.
    $f_0(x) = \underbrace{\text{argmin}}_{c} \sum_{i=1}^{N} L(y_i, c)$
2. For iterations $t$ = 1 to $T$ do:
(a) For samples $i$ = 1 to $N$ do:
    $r_{ti} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}$
  End for
  (b) Use $(x_i, r_{ti})$ $(i = 1,\ 2, \ldots, N)$ to fit the regression tree. The leaf node region is $R_{tj}, j = 1,\ 2, \ldots,\ J$, where $J$ is the number of leaf nodes.
  (c) For $j$ = 1 to $J$ do:
    $c_{tj} = \underbrace{\text{argmin}}_{c} \sum_{x_i \in R_{tj}} L(y_i, f_{t-1}(x_i) + c)$
  End for
(d) Update
    $f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I\left(x \in R_{tj}\right)$
End for
3. Return Output strong learner
    $f(x) = f_T(x) = \sum_{t=1}^{T} \sum_{j=1}^{J} c_{tj} I\left(x \in R_{tj}\right)$
```

5. Performance Evaluation

5.1. Data Set

To ensure the authenticity of data, the public data set, i.e., the patient clinical data and human body's health indicators (Tables 2 and 3), used for indirect prediction of blood glucose (fasting/pre-prandial) levels in the experiments presented here, were provided by a tertiary-care hospital in 2017 as part of the Tianchi competition [43]. The hospital keeps the physical examination results of each patient and files the data. The patients' names were not released to protect their privacy; these were replaced with IDs. A total of 6641 data entries were made publicly available with 42 features. As it is empirically known that the 'patient ID' and 'date of physical examination' data features have no effect on the predicted blood glucose values, these two features were omitted, and the remaining 40 features only were used for training of models, considered in this paper. However, some of the features contain missing values, as shown in Figure 2.

Table 2. Human body's health indicators, divided into groups.

Group	Human Body's Health Indicators
Liver functioning	Aspartate aminotransferase, Alanine aminotransferase, Alkaline phosphatase, γ-Glutamyltransferase, Total serum protein, Serum albumin, Globulin, Ratio of albumin to globulin.
Blood fat	Serum triglyceride, Total cholesterol in lipoproteins, High-density lipoprotein cholesterol, Low-density lipoprotein cholesterol.
Kidney functioning	Urea, Creatinine, Uric acid.
Hepatitis B virus	Hepatitis B surface antigen, Hepatitis B surface antibody, Hepatitis Be antigen, Hepatitis Be antibody, Hepatitis B core antibody.
Blood routine examination	White blood cell count, Red blood cell count, Hemoglobin, Packed cell volume, Mean corpuscular volume, Mean corpuscular hemoglobin, Mean corpuscular hemoglobin concentration, Red blood cell volume distribution width, Platelet count, Mean platelet volume, Platelet volume distribution width, Platelet specific volume, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils.

Table 3. Human body's health indicators with their normal ranges of values.

Short Name	Full Name/ Explanation	Normal Range of Values	Short Name	Full Name/ Explanation	Normal Range of Values
Age	Patient's age	3 ÷ 93 years	HbeAg	Hepatitis Beantigen	0.0 ÷ 0.5 PEI/mL
Sex	Patient's sex	Male/ Female	HBeAb	Hepatitis Beantibody	0.0 ÷ 1.5 PEI/mL
HBsAg	Hepatitis Bsurface antigen	0.0 ÷ 0.5 ng/mL	HBcAb	Hepatitis B core antibody	0.0 ÷ 0.9 PEI/mL
HBsAb	Hepatitis B surface antibody	0 ÷ 10 miu/mL	WBC	White blood cell count	3.50 ÷ 9.50×10ˆ9/L

Table 3. *Cont.*

Short Name	Full Name/ Explanation	Normal Range of Values	Short Name	Full Name/ Explanation	Normal Range of Values
AST	Aspartateamino-transferase	0 ÷ 40 U/L	RBC	Red blood cell count	4.30 ÷ 5.80×10^12/L
ALT	Alanine aminotransferase	0 ÷ 40 U/L	HGB	Hemoglobin	130 ÷ 175 g/L
ALP	Alkalinephosphatase	35 ÷ 135 U/L	PCV	Packed cell volume	40 ÷ 50%
GGT	γ-Glutamyltransferase	11 ÷ 60 U/L	MCV	Mean corpuscular volume	82 ÷ 100 fL
TP	Total serum protein	60 ÷ 83 g/L	MCH	Mean corpuscular hemoglobin	34 ÷ 37 pg
ALB	Serum albumin	37 ÷ 53 g/L	MCHC	Mean corpuscular hemoglobin concentration	316 ÷ 354 g/L
GLB	Globulin	15.0 ÷ 35.0 g/L	ROW	Red blood cell volume distribution width	9.0 ÷ 17.0 fL
AG	Ratio of albumin to globulin	1.1 ÷ 2.5	PLT	Platelet count	125 ÷ 350×10^9/L
TG	Serumtriglyceride	0.00 ÷ 1.71 mmol/L	MPV	Mean platelet volume	6.5–12.0 fL
TC	Total cholesterol in lipoproteins	3.1 ÷ 6.1 mmol/L	PDW	Platelet volume distribution width	9.0 ÷ 17.0 fL
HDL_C	High-density lipoprotein cholesterol	0.9 ÷ 2.0 mg/dL	PCT	Platelet specific volume	0.108 ÷ 0.282%
LDL_C	Low-density lipoprotein cholesterol	0.00 ÷ 3.36 mg/dL	Neutrophils	Percentage of neutrophilsin WBC	40.0 ÷ 75.0%
Urea	Urea	2.82 ÷ 8.20 mmol/L	Lymph	Percentage of lymphocytes in WBC	20.0 ÷ 50.0%
Cre	Creatinine	45 ÷ 106 µmol/L	Monocytes	Percentage of monocytes in WBC	3.0 ÷ 10.0%
UA	Uric acid	200 ÷ 450 µmol/L	Eosinophils	Percentage of eosinophils in WBC	0.4 ÷ 8.0%
bgl	Blood glucose (fasting/pre-prandial) level	4.0 ÷ 6.0 mmol/L	Basophils	Percentage of basophils in WBC	0.0 ÷ 1.0%

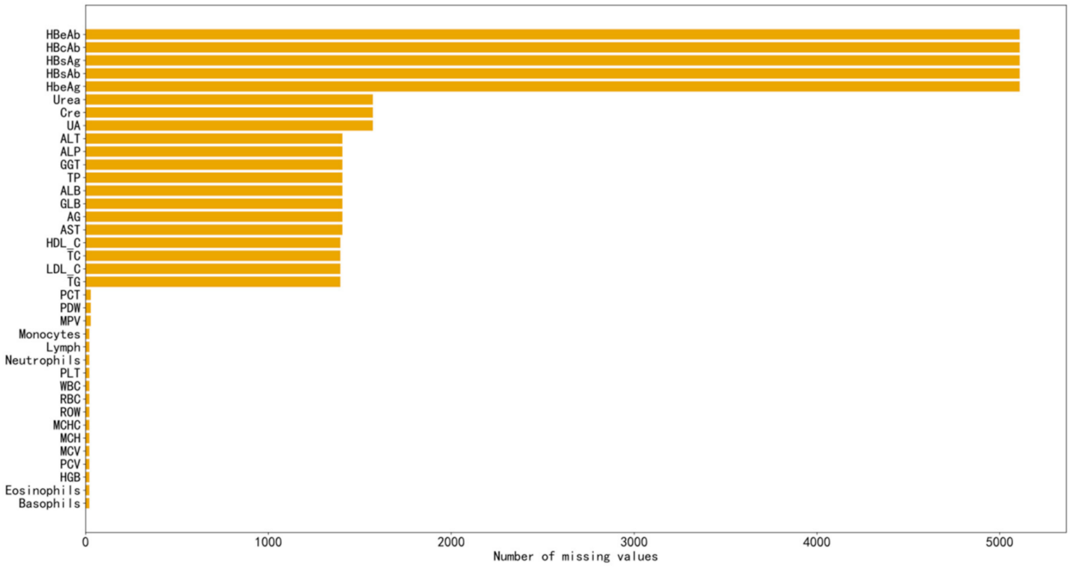

Figure 2. The number of missing values per data feature.

The influence weight of each data feature was obtained according to a correlation function. The importance degree of each data feature is depicted on Figure 3.

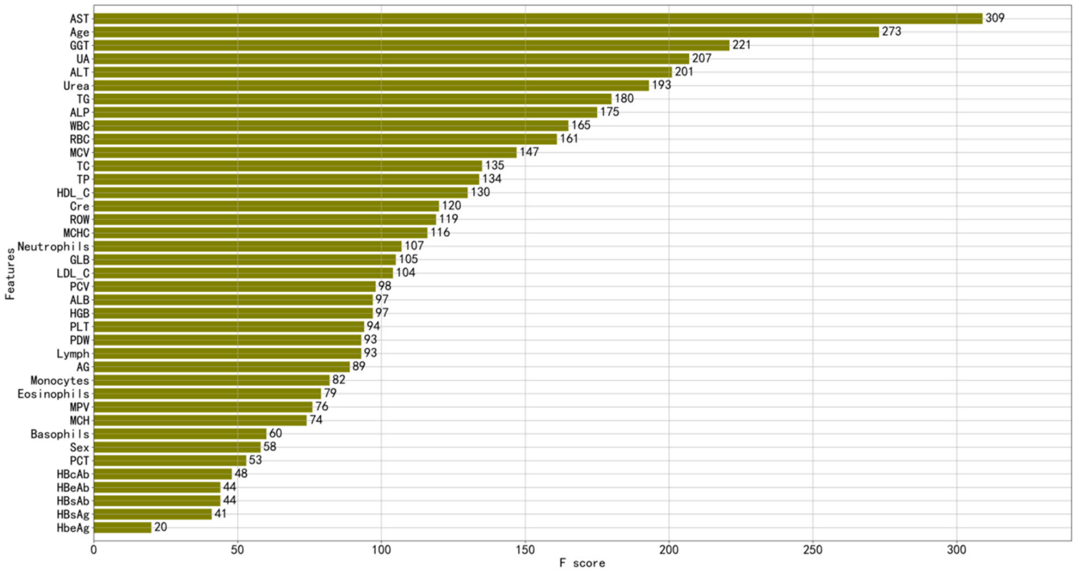

Figure 3. Eigenvalue weight.

From Figure 2, one can see that the top five data features with severely missing values are HBeAb, HBcAb, HBsAg, HBsAb, and HbeAg. Even though the eigenvalue weight of these five features (Figure 3) is small, in a clinical sense these features have a certain impact on the blood glucose levels. Thus, in order to avoid wasting the information contained in these data features, differently from [11], these were not just deleted but rather filled with the medians. The results obtained from the experiments, described in the next subsection, confirmed that this tactic works better than simply deleting all data features with severely missing values.

Due to measurement equipment's problems or presence of outliers in some attributes of human factors, in order to avoid poor prediction based on the normal data due to fitting the outliers, we used the boxplots to display the outliers in each attribute, by setting the outlier empty and filling it and the other missing data simultaneously with the medians. Figure 4 shows the boxplot of γ-Glutamyltransferase.

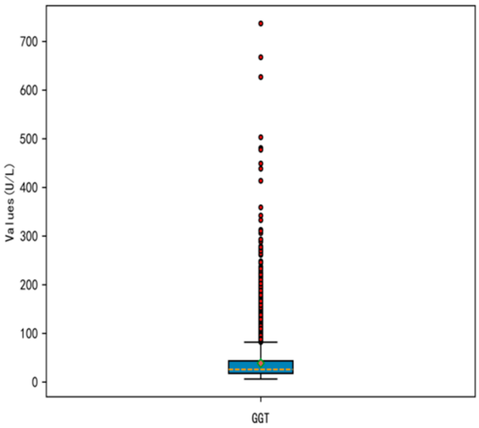

Figure 4. Boxplot of γ-Glutamyltransferase.

5.2. Experiments

The training set $\{x_1, x_2, \ldots, x_N\}$, used in the experiments, comprised $N = 5641$ samples, whereas the test set $\{y_1, y_2, \ldots, y_M\}$ consisted of $M = 1000$ samples.

The predicted values of the M samples are as follows:

$$\tilde{y} = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_M\} \tag{8}$$

The experimental process, performed with the considered boosting ensemble learning models enriched by the proposed RG hyperparameter optimization approach, is depicted in Figure 5.

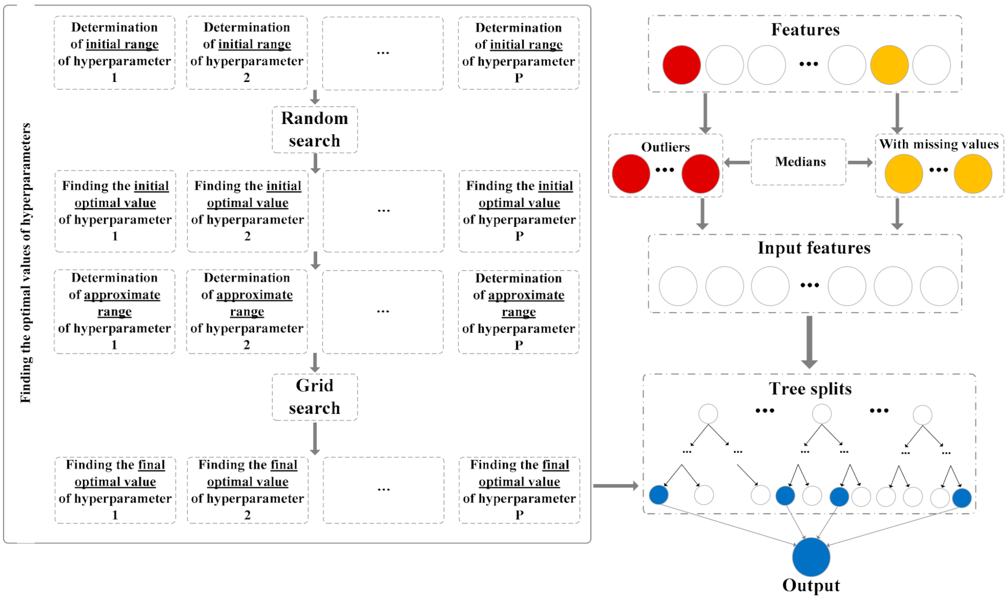
Finding the optimal values of hyperparameters
Determination of initial range of hyperparameter 1
Determination of initial range of hyperparameter 2
...
Determination of initial range of hyperparameter P
Random search
Finding the initial optimal value of hyperparameter 1
Finding the initial optimal value of hyperparameter 2
...
Finding the initial optimal value of hyperparameter P
Determination of approximate range of hyperparameter 1
Determination of approximate range of hyperparameter 2
...
Determination of approximate range of hyperparameter P
Grid search
Finding the final optimal value of hyperparameter 1
Finding the final optimal value of hyperparameter 2
...
Finding the final optimal value of hyperparameter P
Features
Outliers
Medians
With missing values
Input features
Tree splits
Output

Figure 5. The flow chart of the experiments, conducted with the boosting ensemble learning models, enriched by the proposed RG hyperparameter optimization approach.

In the experiments, conducted with the considered boosting ensemble learning models, enriched by the proposed RG hyperparameter optimization approach (further called shortly RG-enriched models), these steps were followed:

1. The data were first divided into a training set and a test set. Then outliers in the data were identified using the IQR method, i.e., all data value less than Q1 − 1.5IQR or greater than Q3 + 1.5IQR were considered outliers. These outliers were set to null value and, along with other missing data, were filled simultaneously with the medians as to avoid the model fitting the outliers and to reduce inaccurate prediction.
2. A random search was then used to find the initial optimal value of each hyperparameter and determine its approximate range. Then, a grid search was performed in this range to find the final optimal value of each hyperparameter, which was brought into the corresponding boosting ensemble learning model (i.e., AdaBoost, GBDT, XGBoost, LightGBM), used for predicting the blood glucose levels.
3. The blood glucose level prediction performance of each RG-enriched model was compared to the corresponding original model by using MSE, root MSE (RMSE), and coefficient of determination R^2 which are commonly used evaluation indicators in regression tasks [11]. The smaller the MSE and RMSE, the better the prediction performance of the corresponding model. For the coefficient of determination: if a model predicts exactly all observed values, then $R^2 = 1$; if a model always predicts the mean of observed values, then $R^2 = 0$; and if a model predicts worse than this, then $R^2 < 0$.

Additionally, as suggested, e.g., in [44,45], an experimental test with randomly generated numbers and a known distribution function was carried out to see if the RG optimization approach does provide improved results to these as well, i.e., to check whether some medical-physiological dependencies are behind the proposed approach. For this purpose, a generated matrix of 3000 rows and 40 columns containing random, uniformly distributed, sample values was used in lieu of real clinical data set. Each column corresponded to one of the 40 human body's health indicators used for the prediction of blood glucose levels in this paper. The values in each of these columns were generated randomly with a uniform distribution, within the relevant ranges shown in Table 3. The values of the 40th column were randomly generated with a uniform distribution within the range of 4.0 to 8.0 as 'phantom values' of blood glucose (measured in mmol/L). A total 2000 out of the 3000 rows were randomly chosen for training, whereas the remaining 1000 rows were used for testing the boosting ensemble learning models considered (i.e., AdaBoost, GBDT, XGBooST, and LightGBM), first in their original form and then by applying the RG hyperparameter optimization to see if this could bring any improvement in predicting the values of the last column.

Finally, experiments were performed with the ANN, described in Section 3.4, and the HY_LightGBM model of [11], both applied to the same clinical data set in order to compare their performance to that of the other models considered.

5.3. Results

The results of the first group of experiments, shown in Table 4 and Figure 6, prove that each RG-enriched boosting ensemble learning model outperforms the corresponding original model, according to all evaluation indicators used. In terms of MSE and RMSE, for instance, the biggest improvement is achieved against the XGBoost model (24.40% for MSE and 11.54% for RMSE) and the smallest improvement against the GBDT model (1.47% for MSE and 0.75% for RMSE).

Table 4. The blood glucose level prediction improvement, in terms of MSE, RMSE and R^2, of RG-enriched boosting ensemble learning models against corresponding original models.

Model	**MSE**	**RMSE**	**R2**
RG_AdaBoost	*0.3766* *(19.07% improvement)*	*0.6137* *(9.11% improvement)*	*0.1195*
AdaBoost	0.4484	0.6696	−0.0481
RG_GBDT	*0.3741* *(1.47% improvement)*	*0.6116* *(0.75% improvement)*	*0.1255*
GBDT	0.3797	0.6162	0.1123
RG_XGBoost	*0.3787* *(24.40% improvement)*	*0.6154* *(11.54% improvement)*	*0.1146*
XGBoost	0.4711	0.6864	−0.1012
RG_LightGBM	*0.3871* *(6.90% improvement)*	*0.6222* *(3.39% improvement)*	*0.0950*
LightGBM	0.4138	0.6433	0.0326
HY_LightGBM *	0.4135	0.6430	0.0333
ANN	0.5180	0.7200	−0.2264

* The figures for the HY_LightGBM model differ from the figures reported in [11], as these were obtained by us based on our own implementation of this model (in Python) and applying it on the same data set, as in [11], but using only the publicly available part of it, totaling in 6641 data entries, and excluding the non-publicly available 1001 data entries used in [11].

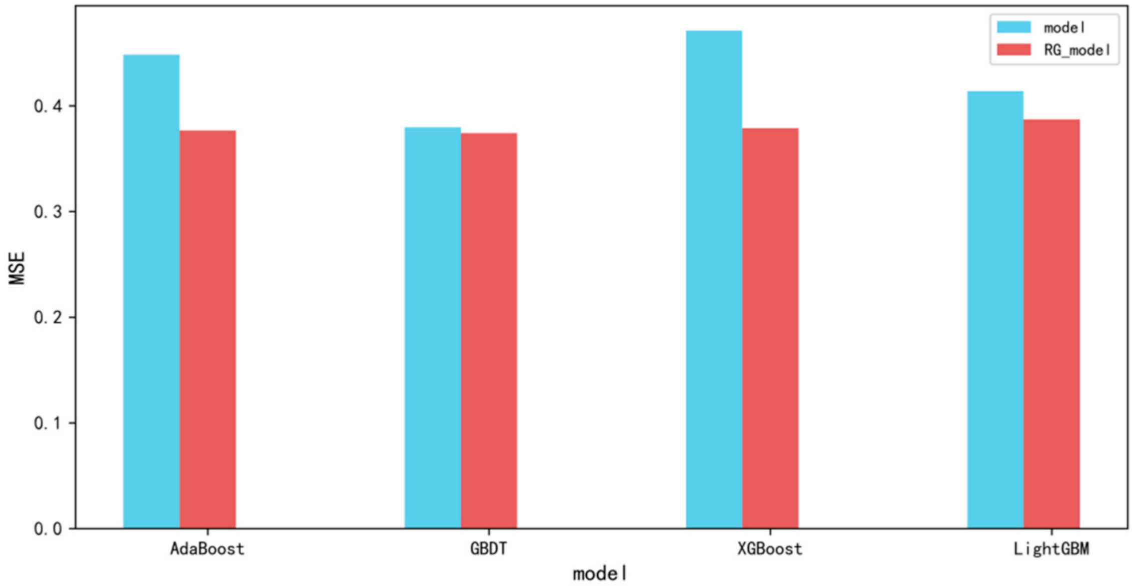

Figure 6. The prediction performance improvement, in terms of MSE, of boosting ensemble learning models after enriching them by the proposed RG hyperparameter optimization.

Figure 7 shows the difference (d) between the predicted and actual blood glucose (fasting/pre-prandial) levels for the four RG-enriched models. Thanks to the good prediction ability of the proposed models, most of the absolute values of d are less than 1 mmol/L; a few of these are in the range of 1 to 2 mmol/L, and only a very few are in the range of 2 to 3 mmol/L. Therefore, based on the specified diagnostic ranges of blood glucose (fasting/pre-prandial) levels [46], it can be considered that most of the errors produced by the proposed models are within the acceptable margin separating the two groups of people—without and with type 2 diabetes, i.e., 4.0 ÷ 6.0 mmol/L and over 7.0 mmol/L, respectively.

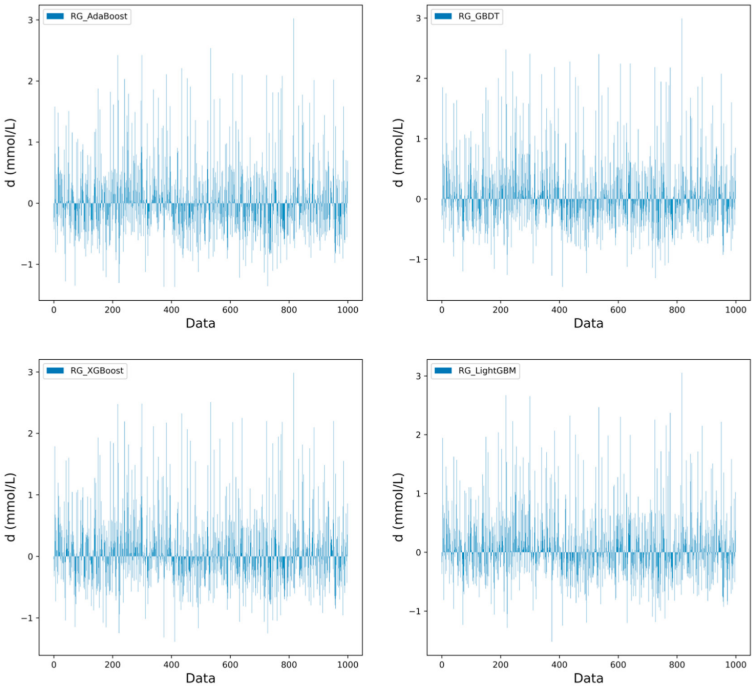

Figure 7. The difference (d) between the predicted and actual blood glucose (fasting/pre-prandial) levels for the RG-enriched models.

Figure 8 depicts the Bland–Altman plots for the four RG-enriched models. As significant part of the data points falls within ±1.96 standard deviations (SD) of the mean difference (MD), the blood glucose values predicted by the RG-enriched model are in good agreement with the actual values [47].

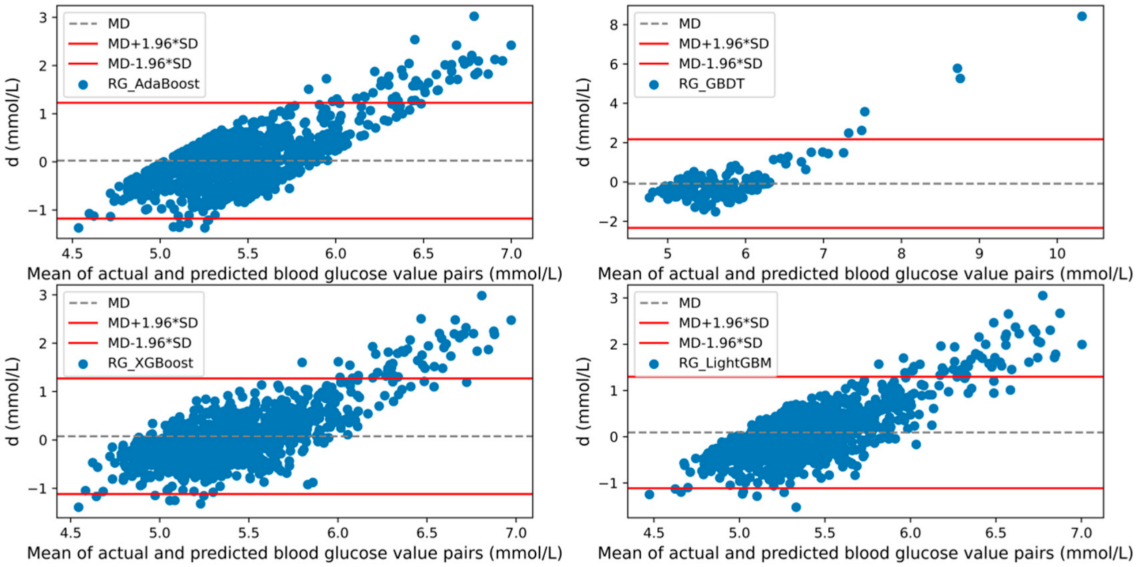

Figure 8. The Bland–Altman plots for the RG-enriched models.

The scatter plots and Clarke error grid analysis (EGA) diagrams of the actual blood glucose levels versus the levels predicted by the four RG-enriched models are depicted in Figures 9–17. As evident from the EGA diagrams, for all RG-enriched models, a dominant part of values is in Zone A representing accurate blood glucose prediction results, some other values are in Zone B representing acceptable prediction results, and only a small percentage of values are in Zone D representing failure to detect and treat diabetes [48].

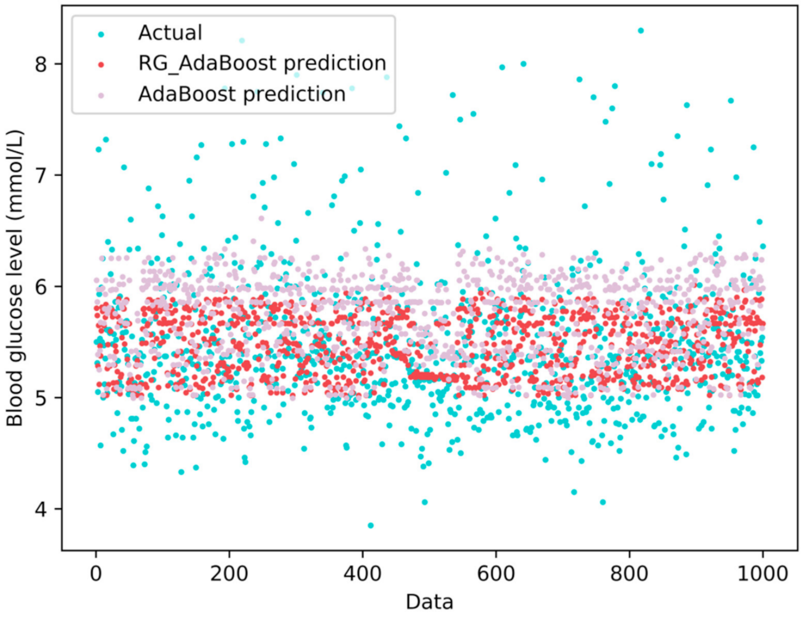

Figure 9. The scatter plot diagram of the actual and predicted blood glucose (fasting/pre-prandial) levels by the RG_AdaBoost and AdaBoost models.

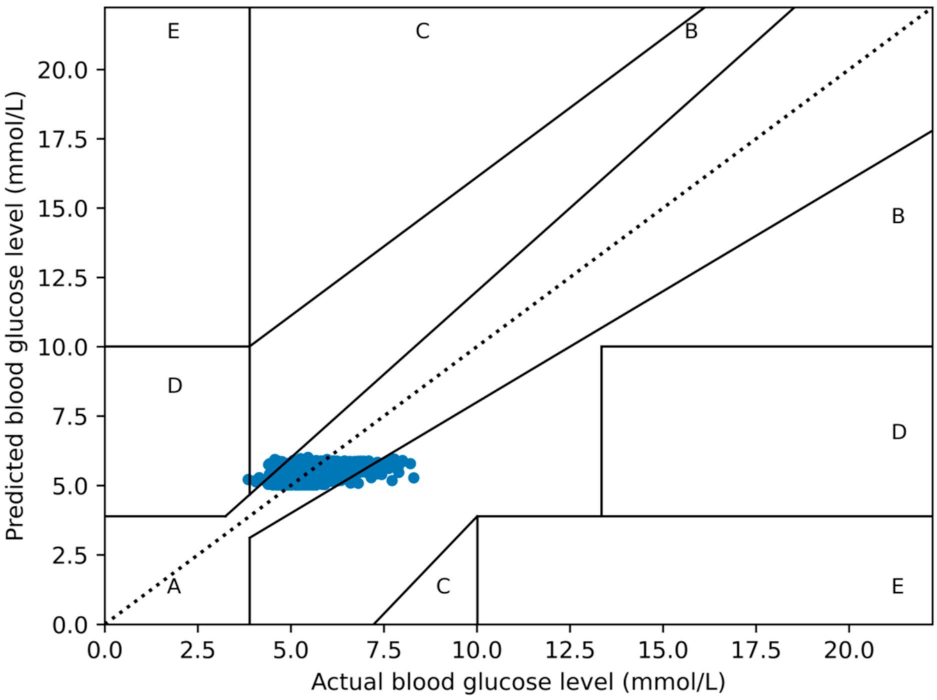

Figure 10. The EGA diagram for the RG_AdaBoost model.

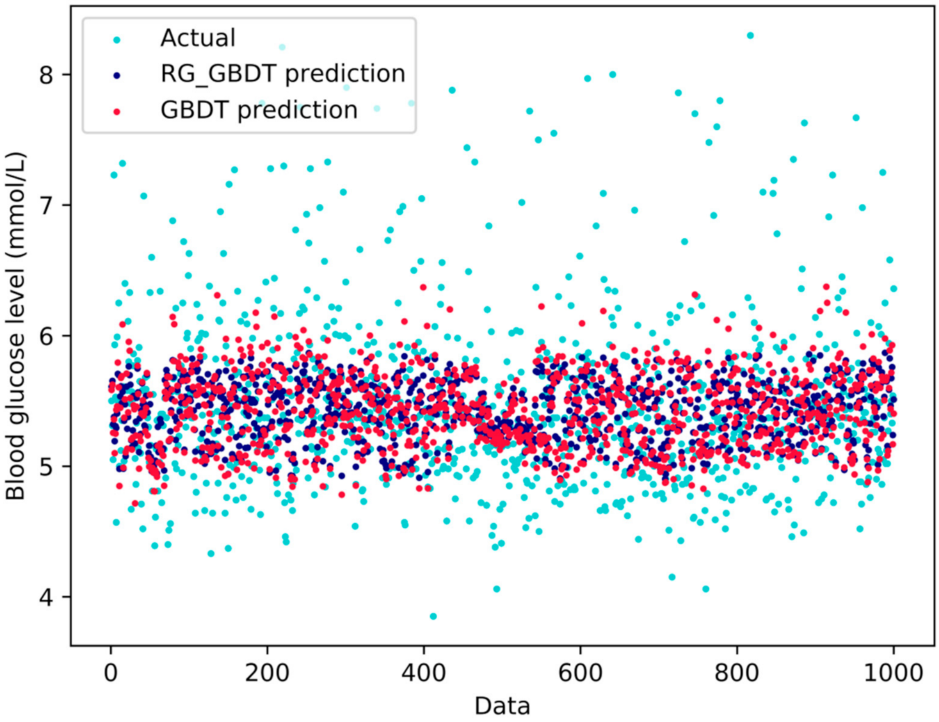

Figure 11. The scatter plot diagram of the actual and predicted blood glucose (fasting/pre-prandial) levels by the RG_GBDT and GBDT models.

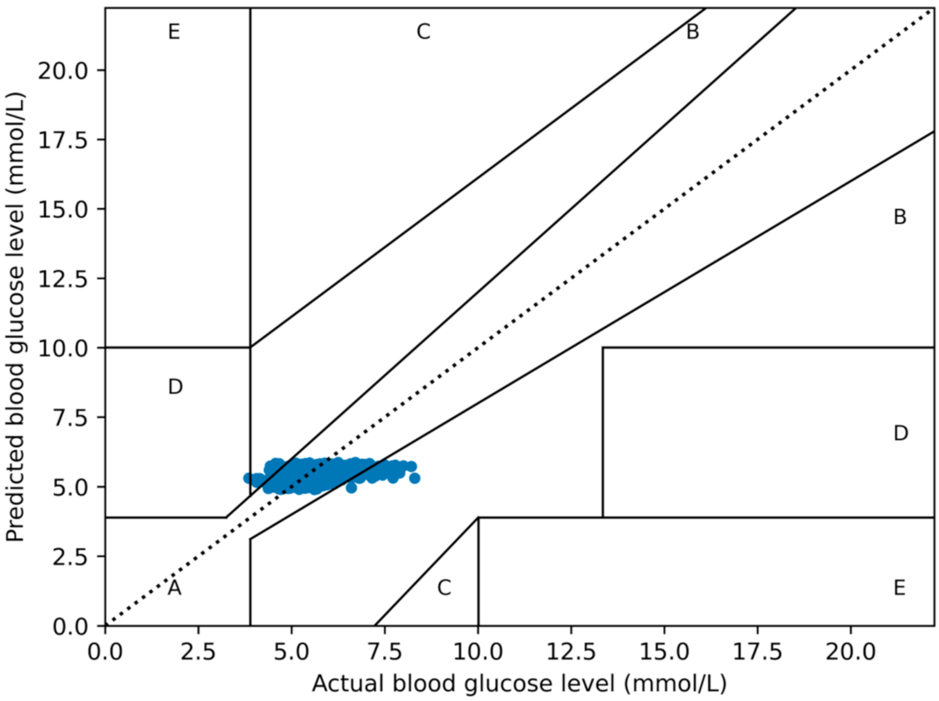

Figure 12. The EGA diagram for the RG_GBDT model.

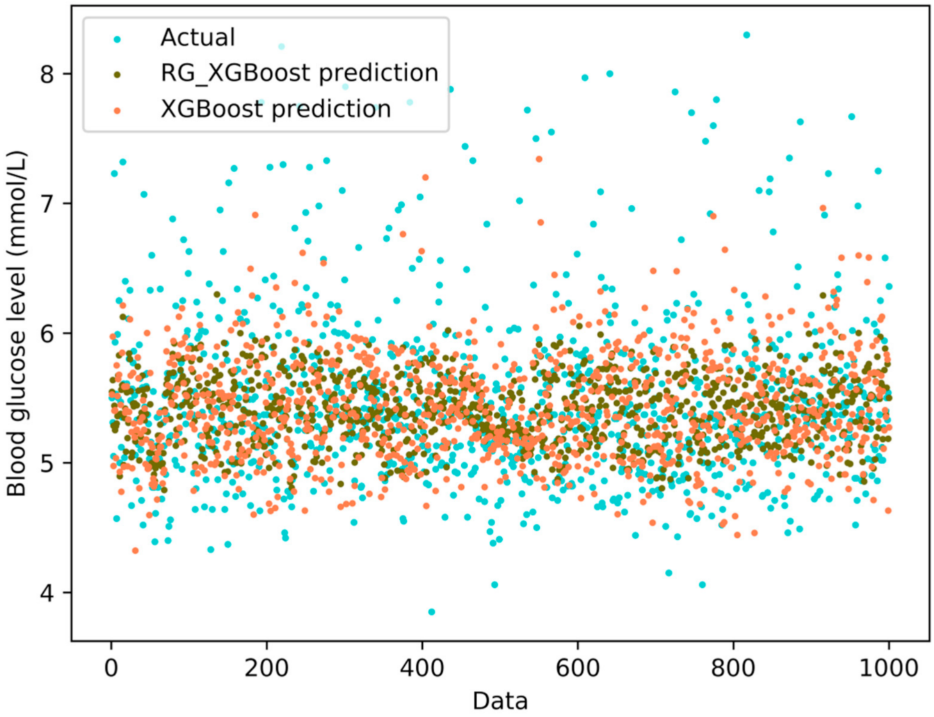

Figure 13. The scatter plot diagram of the actual and predicted blood glucose (fasting/pre-prandial) levels by the RG_XGBoost and XGBoost models.

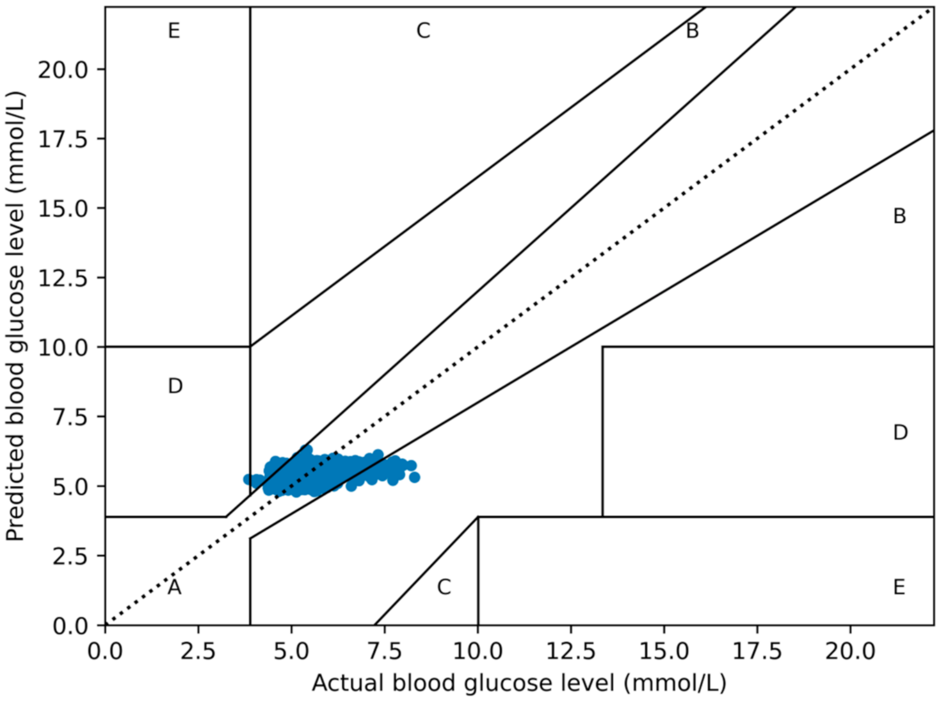

Figure 14. The EGA diagram for the RG_XGBoost model.

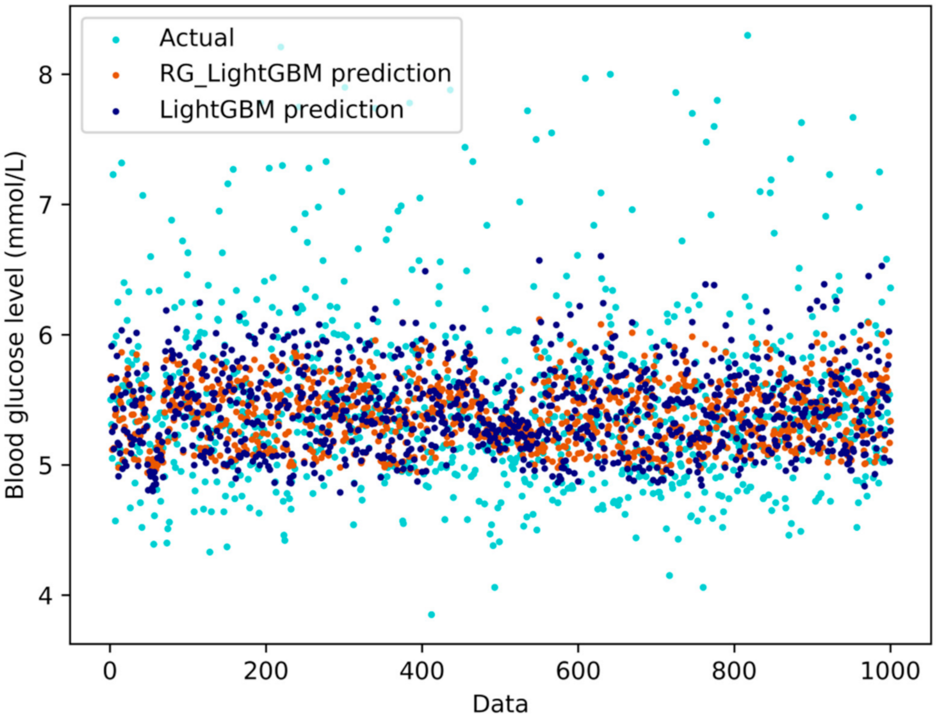

Figure 15. The scatter plot diagram of the actual and predicted blood glucose (fasting/pre-prandial) levels by the RG_LightGBM and LightGBM models.

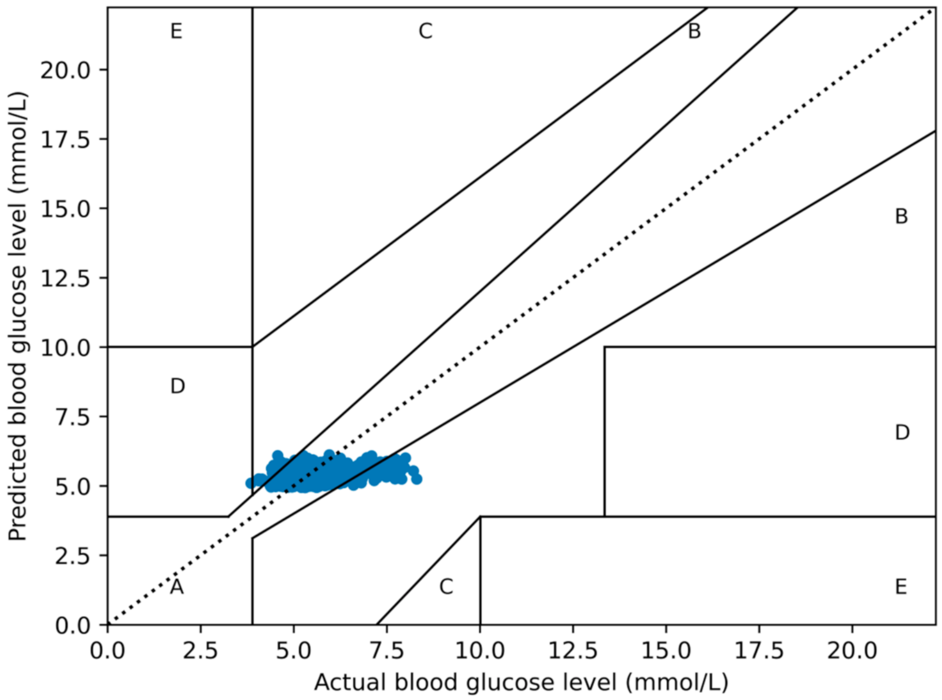

Figure 16. The EGA diagram for the RG_LightGBM model.

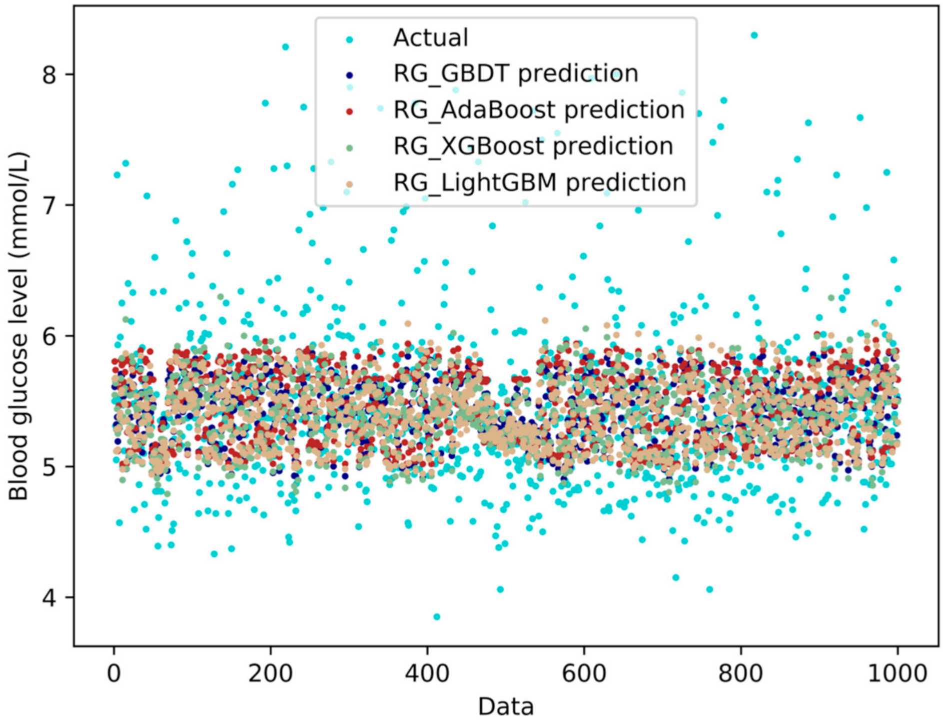

Figure 17. The scatter plot diagram of the actual and predicted blood glucose (fasting/pre-prandial) levels by the RG-enriched models.

The results of the last group of experiments (c.f. the last two lines in Table 4) show that: (i) each of the proposed RG-enriched boosting ensemble learning models outperforms the HY_LightGBM model [11], even though the latter performs better than three of the original boosting ensemble learning models considered, i.e., AdaBoost, XGBoost, and LightGBM; and (ii) the artificial neural network (ANN) performs worst, even worse than all the original boosting ensemble learning models.

The results, shown in Table 5, correspond to the experimental test with random numbers and a known distribution function, which was carried out to check if the proposed RG optimization approach does provide any improved results to these as well. The obtained results demonstrate that: (i) the prediction performance of the boosting ensemble learning models is worse when applying them on random uncorrelated data rather than on real clinical data, as evident from the MSE, RMSE, and R^2 values shown in Table 5, which are all worse than the corresponding ones shown in Table 4; and (ii) in the case of random data, the original boosting ensemble models perform better than their RG-form, which proves that the proposed RG hyperparameter optimization approach works only on real correlated clinical data, used for indirect prediction of blood glucose levels.

Table 5. The negative effect of the RG hyperparameter optimization on the prediction performance of boosting ensemble learning models when applied on randomly generated uncorrelated data.

Model	MSE	RMSE	R2
RG_AdaBoost	*1.3701* *(0.15% deterioration)*	*1.1705* *(0.08% deterioration)*	*−0.0107*
AdaBoost	1.3681	1.1697	−0.0092
RG_GBDT	*1.4190* *(1.08% deterioration)*	*1.1912* *(0.54% deterioration)*	*−0.0468*
GBDT	1.4039	1.1849	−0.0357
RG_XGBoost	*1.7010* *(2.16% deterioration)*	*1.3042* *(1.07% deterioration)*	*−0.2548*
XGBoost	1.6651	1.2904	−0.2284
RG_LightGBM	*1.5219* *(0.57% deterioration)*	*1.2337* *(0.28% deterioration)*	*−0.1227*
LightGBM	1.5133	1.2302	−0.1163

6. Conclusions

It is well known that ensemble learning can lead to better prediction results compared to regular machine learning based on a single model. However, as the selection of different hyperparameters has a great impact on the prediction results, this should be done with caution. In order to improve the prediction performance of boosting ensemble learning models, this paper has proposed to enrich these by an RG hyperparameter optimization, involving a sequential use of a random search (R) and a grid search (G). Based on this RG double optimization, the prediction performance of the considered state-of-the-art boosting ensemble learning models has been improved (significantly in some cases) as demonstrated by the conducted experiments for predicting blood glucose levels in patients, based on their clinical data.

Considering that a small error in medicine can cause an immense damage to patients and hospitals, it is clear that every bit of improvement in predicting the patients' health condition is important. The RG hyperparameter optimization approach, proposed in this paper, could be helpful in increasing the work efficiency and accuracy of healthcare providers and in supporting intelligent medical treatment. As such, it shows a great promise for use in clinical applications and is worthy of further study.

The future work in this direction will be focused on: (1) using the mean values, instead of the median values, for filling the missing data; (2) performing principal component analysis for better feature selection; (3) Bayesian optimization fused with the proposed RG approach for the purpose of exploring the pathogenesis of diabetes; and (4) considering more human body's health indicators having an impact on the blood glucose level.

Author Contributions: Conceptualization, Y.W. and Z.J.; methodology, Y.W.; validation, I.G., H.Z.; formal analysis, H.Z.; writing—original draft preparation, Y.W.; writing—review and editing, I.G.; supervision, Y.A.; project administration, Z.J. All authors have read and agreed to the published version of the manuscript.

Funding: This publication has emanated from a joint research conducted with the financial support of the National Key Research and Development Program of China under the Grant No. 2017YFE0135700 and the Bulgarian National Science Fund (BNSF) under the Grant No. КП-06-ИП-КИТАЙ/1 (KP-06-IP-CHINA/1).

Data Availability Statement: The data are from Alibaba Cloud Labeled Chinese Dataset for diabetes. Available online: https://tianchi.aliyun.com/dataset/dataDetail?dataId=22288 (accessed 29 April 2019).

Acknowledgments: We thank Alibaba for data support.

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

References

1. Saeedi, P.; Salpea, P.; Karuranga, S.; Petersohn, I.; Malanda, B.; Gregg, E.W.; Unwin, N.; Wild, S.H.; Williams, R. Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **2020**, *162*, 108086. [CrossRef]
2. Jahangir, M.; Afzal, H.; Ahmed, M.; Khurshid, K.; Nawaz, R. An expert system for diabetes prediction using auto tuned multi-layer perceptron. In Proceedings of the 2017 Intelligent Systems Conference, London, UK, 7–8 September 2017; IEEE: Piscataway, NJ, USA, 2018.
3. Li, Y.; Teng, D.; Shi, X. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: National cross sectional study. *BMJ* **2020**, *369*, m997. [CrossRef]
4. Iribarren, M.A.; Tejedor, X.; Sanjaume, A.S.; Leis, A.; Doladé Botias, M.; Morales-Indiano, C. Performance evaluation of the new hematology analyzer UniCel DxH 900. *Int. J. Lab. Hematol.* **2021**, 1–9. [CrossRef]
5. Wilbert, V.G.; Ahmed, T.M.; Amin, A. The Progress of Glucose Monitoring—A Review of Invasive to Minimally and Non-Invasive Techniques, Devices and Sensors. *Sensors* **2019**, *19*, 800.
6. Teymourian, H.; Barfidokht, A.; Wang, J. Electrochemical glucose sensors in diabetes management: An updated review (2010–2020). *Chem. Soc. Rev.* **2020**, *49*, 7671–7709. [CrossRef]
7. Klonoff, D.C. Overview of Fluorescence Glucose Sensing: A Technology with a Bright Future. *J. Diabetes Sci. Technol.* **2012**, *6*, 1242–1250. [CrossRef] [PubMed]
8. Malik, B.H.; Cote, G.L. Real-time, closed-loop dual-wavelength optical polarimetry for glucose monitoring. *J. Biomed. Opt.* **2010**, *15*, 017002. [CrossRef] [PubMed]
9. Kubihal, S.; Goyal, A.; Gupta, Y.; Khadgawat, R. Glucose measurement in body fluids: A ready reckoner for clinicians. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**, *15*, 45–53. [CrossRef]
10. Ichai, C.; Preiser, J.C. International recommendations for glucose control in adult non diabetic critically ill patients. *Crit. Care* **2010**, *14*, 166. [CrossRef]
11. Wang, Y.; Wang, T. Application of Improved LightGBM Model in Blood Glucose Prediction. *Appl. Sci.* **2020**, *10*, 3227. [CrossRef]
12. Kim, C.H.; Park, J.Y.; Lee, K.U.; Kim, J.H.; Kim, H.K. Association of Serum γ-Glutamyl Transferase and Alanine Aminotransferase Activities with Risk of Type 2 Diabetes Mellitus Independent of Fatty Liver. *Diabetes Metab. Res. Rev.* **2009**, *25*, 64–69. [CrossRef]
13. Nofal, M.W. Could Liver Functions Predict Type 2 Diabetes Mellitus in Young Obese Men in Najran, Saudi Arabia? *Life Sci. J.* **2013**, *10*, 1498–1503.
14. Zhao, W.; Tong, J.; Liu, J.; Liu, J.; Li, J.; Cao, Y. The Dose-Response Relationship between Gamma-Glutamyl Transferase and Risk of Diabetes Mellitus Using Publicly Available Data: A Longitudinal Study in Japan. *Int. J. Endocrinol.* **2020**, *2020*, 1–7. [CrossRef]
15. Ahn, H.R.; Shin, M.H.; Nam, H.S.; Park, K.S.; Lee, Y.H.; Jeong, S.K.; Choi, J.S.; Kweon, S.S. The association between liver enzymes and risk of type 2 diabetes: The Namwon study. *Diabetol. Metab. Syndr.* **2014**, *6*, 14. [CrossRef]
16. Wu, M.; Nian, S.; Feng, L.; Bai, X.; Ye, D.; Zhang, C.; Yan, Z.; Ma, Q.; Shao, C.; Bi, Q.; et al. Potential role of liver enzymes for predicting elevated blood glucose levels. *Can. J. Physiol. Pharmacol.* **2021**. [CrossRef]
17. Wei, M.; Gibbons, L.W.; Mitchell, T.L.; Kampert, J.B.; Lee, C.D.; Blair, S.N. The Association between Cardiorespiratory Fitness and Impaired Fasting Glucose and Type 2 Diabetes Mellitus in Men. *Ann. Intern. Med.* **1999**, *130*, 89–96. [CrossRef]
18. Bos, G.; Dekker, J.M.; Nijpels, G.; De Vegt, F.; Diamant, M.; Stehouwer, C.D.A.; Bouter, L.M.; Heine, R.J. A Combination of High Concentrations of Serum Triglyceride and Non-High-Density-Lipoprotein-Cholesterol is a Risk Factor for Cardiovascular Disease in Subjects with Abnormal Glucose Metabolism-The Hoorn Study. *Diabetologia* **2003**, *46*, 910–916. [CrossRef] [PubMed]
19. Zoppini, G.; Targher, G.; Negri, C.; Stoico, V.; Gemma, M.L.; Bonora, E. Usefulness of the Triglyceride to High-Density Lipoprotein Cholesterol Ratio for Predicting Mortality Risk in Type 2 Diabetes: Role of Kidney Dysfunction. *Atherosclerosis* **2010**, *212*, 287–291. [CrossRef] [PubMed]
20. Higuchi, S.; Izquierdo, M.C.; Haeusler, R.A. Unexplained Reciprocal Regulation of Diabetes and Lipoproteins. *Curr. Opin. Lipidol.* **2018**, *29*, 186–193. [CrossRef]
21. Chhatriwala, M.N.; Patel, M.P.; Patel, D.S.; Shah, H.N. Relationship between Dyslipidemia and Glycemic Status in Type-2 Diabetes Mellitus. *Natl. J. Lab. Med.* **2019**, *8*. [CrossRef]

22. Palaniappan, S.; Awang, R. Intelligent heart disease prediction system using data mining techniques. In Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 31 March–4 April 2008; IEEE: Piscataway, NJ, USA, 2008.
23. Vijayarani, S.; Dhayanand, S. Liver disease prediction using SVM and Naïve Bayes algorithms. *Int. J. Sci. Eng. Technol. Res.* **2015**, *4*, 816–882.
24. Solanki, Y.S.; Chakrabarti, P.; Jasinski, M.; Leonowicz, Z.; Bolshev, V.; Vinogradov, A.; Jasinska, E.; Gono, R.; Nami, M. A Hybrid Supervised Machine Learning Classifier System for Breast Cancer Prognosis Using Feature Selection and Data Imbalance Handling Approaches. *Electronics* **2021**, *10*, 699. [CrossRef]
25. Santhanam, T.; Padmavathi, M.S. Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. *Procedia Comput. Sci.* **2015**, *47*, 76–83. [CrossRef]
26. Nai-arun, N.; Sittidech, P. Ensemble Learning Model for Diabetes Classification. *Adv. Mater. Res.* **2014**, *931–932*, 1427–1431. [CrossRef]
27. Chen, T.Q.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016; pp. 785–794.
28. Ke, G.L.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2017; pp. 3146–3154.
29. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
30. Bergstra, J.; Bengio, Y. Random Search for HyperParameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
31. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
32. Zhou, Z.H. *Ensemble Learning*, 2nd ed.; Li, S.Z., Ed.; Springer: Boston, MA, USA, 2009.
33. Schapire, R.E. The strength of weak learnability. *Mach. Learn.* **1990**, *5*, 197–227. [CrossRef]
34. Tu, J.V. Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. *J. Clin. Epidemiol.* **1996**, *49*, 1225–1231. [CrossRef]
35. Xia, S.Y.; Xia, Y.L.; Yu, H.; Liu, Q.; Luo, Y.; Wang, G.; Chen, Z. Transferring Ensemble Representations Using Deep Convolutional Neural Networks for Small-Scale Image Classification. *IEEE Access* **2019**, *7*, 168175–168186. [CrossRef]
36. Kingma, D.; Ba, J.L. ADAM: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015. m997.
37. Bühlmann, P.; Yu, B. Boosting with the L2-Loss: Regression and Classification. *J. Am. Stat. Assoc.* **2003**, *98*, 324–339. [CrossRef]
38. Gao, X.L.; Fang, Y.X. A note on the generalized degrees of freedom under the L1 loss function. *J. Stat. Plan. Inference* **2011**, *141*, 677–686. [CrossRef]
39. Regression Loss Functions All Machine Learners Should Know: Choosing the Right Loss Function for Fitting a Model. Available online: https://heartbeat.fritz.ai/5-regression-loss-functions-all-machine-learners-should-know-4fb140e9d4b0 (accessed on 20 March 2021).
40. Sun, Q.; Zhou, W.; Fan, J. Adaptive Huber Regression. *J. Am. Stat. Assoc.* **2017**, *115*, 254–265, m997. [CrossRef]
41. Al-Fugara, A.K.; Ahmadlou, M.; Al-shabeeb, A.R.; AlAyyash, S.; Al-Amoush, H.; Al-Adamat, R. Spatial mapping of groundwater springs potentiality using grid search-based and genetic algorithm-based support vector regression. *Geocarto Int.* **2020**, *35*, 1–22. [CrossRef]
42. Liu, X.; Tan, W.; Tang, S. A Bagging-GBDT ensemble learning model for city air pollutant concentration prediction. In Proceedings of the 4th International Conference on Advances in Energy Resources and Environment Engineering, Chengdu, China, 7–9 December 2018.
43. Alibaba Cloud Labeled Chinese Dataset for Diabetes. Available online: https://tianchi.aliyun.com/dataset/dataDetail?dataId=22288 (accessed on 29 April 2019).
44. Arnold, M.A.; Burmeister, J.J.; Small, G.W. Phantom glucose calibration models from simulated noninvasive human near-infrared spectra. *Anal. Chem.* **1998**, *70*, 1773–1781. [CrossRef]
45. Arnold, M.A.; Small, G.W. Noninvasive glucose sensing. *Anal. Chem.* **2005**, *77*, 5429–5439. [CrossRef]
46. Diabetes Meal Plans. Available online: https://diabetesmealplans.com/5080/diabetes-blood-sugar-levels-chart-printable/ (accessed on 13 July 2021).
47. Bland, J.M.; Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1986**, *1*, 307–310. [CrossRef]
48. Clarke, W.L. The Original Clarke Error Grid Analysis (EGA). *Diabetes Technol. Ther.* **2005**, *7*, 776–779. [CrossRef] [PubMed]