# Fusion of multi-regression models based on the histogram information for blood glucose estimation

**Yiting Wei**
Guangdong University of Technology
**Weizhi Guo**
Guangdong University of Technology
**Bingo Wing-Kuen Ling** ( yongquanling@gdut.edu.cn )
Guangdong University of Technology
**Yuheng Dai**
Guangdong University of Technology
**Qing Liu**
Guangdong University of Technology

**Research Article**

**Keywords:** Blood glucose estimation, multi-regression models, fusion, histogram information

**Posted Date:** January 24th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2489524/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Signal, Image and Video Processing on May 16th, 2023. See the published version at https://doi.org/10.1007/s11760-023-02608-w.

# Abstract

Diabetes is a chronic disease that severely degrades the human health. Hence, the blood glucose estimation plays an important role for monitoring the diabetic condition. In order to better estimate the blood glucose values, the multi-regression models are employed. It is worth noting that increasing the total number of the regression models would decrease the regression error. Therefore, this paper proposes a method for fusing the various regression models together based on the histogram information of the blood glucose values in the training set. The computer numerical simulation results show that the regression error yielded by our proposed method is significantly lower than those yielded by the existing methods. Also, our proposed method is also applicable for other regression applications.


# 1 Introduction

Diabetes is a disease resulted to the abnormal blood glucose values caused by the dysfunction or the loss of the insulin secretion of the human pancreas. The long term hyperglycemia or the blood glucose fluctuations can lead to many serious complications including the cardiovascular disease, the diabetic foot, the pathological changes in the eye tissue as well as the structural and functional damage of the kidneys. These complications may also lead to the mental harm to the patients [2]. It is the third most serious chronic disease after the tumor and the cardiovascular disease that threatens the human health. The 2021 IDF Diabetes Atlas (10th Edition) stated that 537 millions of the adults in the worldwide had the diabetes in 2021, in which 6.7 millions of them were died from the diabetes or its complications. This mortality contributed 12.2% of the mortality due to all the reasons. Overall, an average of 1 person was died due to the diabetes or its complications in every 5 seconds [1]. According to the report issued by the IDF Diabetes Atlas (10th Edition), the total number of the adults in the worldwide having the diabetes will reach 783 millions at 2045. At that time, there will be one eighth of the adults in the worldwide having the diabetes [1]. It is worth noting that this corresponds to the 46% of the increment relative to the corresponding figure at 2021. Since the estimated population growth at the same period of time is just 20%, the rapid increase of the adults in the worldwide having the diabetes will introduce a lot of the financial and medical burdens to the government.

However, about 240 millions of the adults having the diabetes in the worldwide are undiagnosed [1]. This corresponds to 44.7% of the whole diabetic population. Since the early symptoms of the diabetes are not obvious, many patients do not have the timely physical examination or the screening. This results to the delays of the diagnosis of the diabetes and the occurrence of the irreversible complications. It is worth noting that the medical officers perform the diagnosis of the diabetes mainly by detecting whether the blood glucose values are within the normal range or not. Nevertheless, the blood glucose values are highly dependent on the amount of intaken carbohydrates and the duration between the measurement and the meal time. However, the eating habits of different individuals are very different. Hence, the detection of the abnormal health conditions of the diabetic patients only via monitoring their blood glucose values is not enough.

Besides, only a limited amount of the blood glucose data is available from an individual diabetic patent and this limited amount of the blood glucose data provided by them may not reflect the real situation of their diabetic conditions. The American Diabetes Association and the other associations have conducted a ten year study on the intensive treatment for the type 1 diabetes. This intensive treatment is with the long term monitoring of the blood glucose values. Compared to the conventional therapy, the results showed that the intensive treatment can reduce the ocular complications by 76%, the renal complications by 54% and the neurological complications by 60% [3]. Overall, the long term monitoring of the blood glucose values is critical and essential for the health monitoring of the potential diabetic patients and the diabetic patients.

Since it is necessary to observe the health conditions of the diabetic patients for a period of time and combine the multiple blood index data together to perform the diagnosis of the diabetes, the statistical analysis of the various data on the human health indices over a period of time is essential. The conventional statistical analysis includes the study on the conditional probability of having a particular disease for a given value of the human health index. This statistical analysis can help the early detection of the signs as well as the timely intervention and the control of the disease. Hence, the statistical analysis on the various data over a period of time plays an important role in the prevention and the control of the diabetes. Besides, there are many factors affecting the blood glucose values. Nevertheless, these factors are unknown and unexplainable from the medical viewpoints. To find out these factors and establish a model governing the relationship between these factors and the blood glucose values, the machine learning approach is employed. In particular, a set of measurements is first taken from the diabetic patients. These measurements form a training set. Then, a regression model is established using the measurements in the training set. Finally, for a given new measurement in the test set, the regression model is used to estimate the blood glucose values. In fact, many machine learning based models have been developed for performing the blood glucose estimation. For examples, a fusion of the linear regression model and the support vector machine model [4] was developed for performing the blood glucose estimation using the diabetes dataset. Moreover, an autoregressive (ARX) model [5] was proposed for handling the various exogenous inputs. The computer numerical simulation results showed that the algorithm can improve the accuracy of the blood glucose estimation. Furthermore, the XGBoost algorithm based on the ensemble learning [6] was proposed for performing the diabetes prediction. In particular, the model adopted the CART regression tree as the base learner as well as employed the acquired real data to train and test the model. Finally, the main parameters in the XGBoost algorithm were adjusted. Besides, an improved algorithm for performing the feature combination based on the XGBoost approach [7] was proposed. More precisely, the algorithm extracted the numerical features from the textual features of the acquired real data. The computer numerical simulation results showed that the diabetes prediction model established based on the data feature splicing XGBoost algorithm yielded a high prediction accuracy. Also, this algorithm was robust to the acquired data and the required executed time was short. In addition, the deep learning model was proposed for predicting the subcutaneous glucose concentration [8]. This model consisted of several prediction layers. Hence, the prediction accuracy was improved for the majority datasets. Although the above methods can achieve the certain

levels of the estimation accuracies, the obtained results based on the single estimation approach is still very limited.

In fact, the more the individual regression models would result to the lower the estimation error [9]. Therefore, this paper proposes a method for fusing the multi-regression models together based on the histogram information of the blood glucose values in the training set. Our proposed algorithm is evaluated using the Azure open dataset [10]. The computer numerical simulation results show that our proposed method can significantly improve the accuracy of the blood glucose estimation compared to the existing methods. The outline of this paper is as follows. Section 2 presents our proposed method. Section 3 presents the computer numerical simulation results. Finally, the conclusion is drawn in Section 4.

## 2 Our Proposed Method

Figure 1 shows the block diagram of our proposed method. First, ten features are extracted from each measurement. Then, the dataset is divided into the training set, the validation sets and the test set. Next, all the feature vectors are normalized to the unit vectors. Finally, the multi-regression models are fused together based on the histogram information of the blood glucose values in the training set for performing the blood glucose estimation.

# 2.1 Dataset

The diabetes dataset used in this paper is downloaded from the Azure open dataset created in North Carolina State University. This dataset contains 442 measurements.

# 2.2 Feature extraction

10 features including the age, the gender, the body mass index, the mean of the blood pressure value and six features related to the serum are extracted from each measurement. Figure 2 shows the box plot diagram of these features. In particular, the center of the box corresponding to each feature is located at their mean and the range of the box is from its plus and minus one standard deviation. Besides, there is a quantitative measure of the progression of the diabetes one year after the baseline.

# 2.3 Segmentation of dataset

The dataset is divided into the training set, the first validation set, the second validation set and the test set. The ratio of the total number of the feature vectors in these four subsets is approximately equal to 2:1:1:1.

# 2.4 Normalization

Let $\hat{\mathbf{y}}_i$ be the feature vector of the $i^{th}$ measurement in the training set. Then, $\hat{\mathbf{y}}_i$ is normalized to the unit energy vector. Let $g_i = \frac{1}{\sqrt{\hat{\mathbf{y}}_i^T \hat{\mathbf{y}}_i}}$ be the normalized gain. Let $\mathbf{y}_i$ be the normalized vector. That is,

$\mathbf{y}_i = g_i \hat{\mathbf{y}}_i$. Likewise, the feature vectors in the first validation set, the second validation set and the test set are also multiplied by $g_i$ to obtain the normalized feature vectors.

## 2.5 Fusion of the multi-regression models

This paper proposes a method for fusing the multi-regression models together based on the histogram information of the blood glucose values in the training set. Figure 3 shows the procedures of our proposed algorithm. The details of the algorithm are as follows.

Step 1: Fig. 4 shows the histogram of the blood glucose values in the training set. The centers of the second column, the third column, the fourth column, the fifth column and the sixth column in the histogram are found. For this training set, these centers are 68.1mg/dl, 96.8mg/dl, 126mg/dl, 154mg/dl and 183mg/dl.

Step 2: The random forest regression model is established using the training set.

Step 3: The feature vectors in the first validation set are used for performing the blood glucose estimation. Let $(x_i, y_i)$ be the Cartesian coordinate of the pair of the $i^{th}$ reference blood glucose value and the $i^{th}$ estimated blood glucose value in the first validation set. Figure 5 shows these coordinates. At the same time, the ideal coordinates based on the centers of these 5 columns in the histogram of the blood glucose values in the training set found in Step 1 are also plot in the figure. In particular, these 5 ideal coordinates are (68.1, 68.1), (96.8, 96.8), (126, 126), (154, 154) and (183, 183). They are on the straight line with the slope equal to one and passing through the origin as shown as the black dots in Fig. 5.

Step 4: Let $a_1$, $a_2$, $a_3$, $a_4$ and $a_5$ be 5 non-overlapped subsets in the first validation set defined based on these 5 ideal coordinates. In particular, the Euclidean distances between each $(x_i, y_i)$ and these 5 ideal coordinates are computed. The $i^{th}$ feature vector and the $i^{th}$ reference blood glucose value are assigned to one of these 5 subsets using the minimum Euclidean distance rule.

Step 5: An individual random forest regression model is established using each subset of the first validation set. Hence, 5 random forest regression models are established in total.

Step 6: For each feature vector in the second validation set, the blood glucose values are estimated using these 5 random forest regression models established in Step 5. Hence, they are 5 estimated blood glucose values for each feature vector in the second validation set.

Step 7: Let $b_1$, $b_2$, $b_3$, $b_4$ and $b_5$ be 5 non-overlapped subsets in the second validation set based on these 5 random forest regression models established in Step 5. In particular, the Euclidean distances between the reference blood glucose value and these 5 estimated blood glucose values for each measurement in the second validation set are computed. The $i^{th}$ feature vector and the $i^{th}$ reference blood glucose value are assigned to one of these 5 subsets using the minimum Euclidean distance rule.

Step 8: A random forest classification model is established using all the feature vectors in the second validation set and all the classification labels obtained in Step 7. Here, the classification label refers to the index of the regression models established in Step 5.

Step 9: For each feature vector in the test set, the index of the regression models is found using the random forest classification model established in Step 8.

Step 10: For each feature vector in the test set, the blood glucose value is estimated using the regression model defined in Step 5 indexed by Step 9.

## 3 Computer Numerical Simulation Results

# 3.1 Performance metrics

This paper employs the squares of the Pearson's correlation coefficient ($R^2$), the mean absolute error (MAE) and the mean squares error (MSE) as the metrics for evaluating the performance of the blood glucose estimation.

# 3.1.1 $R^2$

The $R^2$ is defined as follow:

$$R^2 = \left( \frac{1}{m-1} \sum_{i=1}^{m} \left( \frac{y_i - \mu}{\sigma} \right) \left( \frac{\tilde{y}_i - \tilde{\mu}}{\tilde{\sigma}} \right) \right)^2$$

1

.

The closer the value of $R^2$ to 1 refers to the higher the correlation between the estimated blood glucose values and the reference blood glucose values.

# 3.1.2 MAE

The MAE is defined as follow:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |\tilde{y}_i - y_i|$$

2

.

The smaller the value of MAE refers to the smaller error of the blood glucose estimation.

# 3.1.3 MSE

The MSE is defined as follow:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i - y_i|^2$$

3

.

The smaller the value of MSE refers to the smaller error of the blood glucose estimation.

# 3.2 Comparisons to the existing results

In order to evaluate the effectiveness of our proposed method, various models including the linear regression based model, the *k* nearest neighbor (kNN) based model, the support vector regression (SVR) based models with the various kernel functions, the decision tree based model, the random forest based model, the gradient boosting decision tree (GBDT) based model and the three layer back propagation neural network (BPNN) based model are employed for performing the comparisons. These models are employed because they are commonly used in the machine learning community. For the linear regression based model, the least squares approach is employed for performing the training. This is because the training problem is convex and the global optimal solution can be found analytically. For the kNN based model, it is worth noting that the pattern recognition problems in the algorithm are the 5 class recognition problems. As the value of *k* is usually chosen as 2 to 3 times of the total number of the classes of the recognition problems, the value of *k* in this paper is chosen as 13. For the SVR based model, since the Gaussian kernel function is characterized by the mean vectors and the covariance matrices, it can achieve a good result for the clustering based dataset. Besides, since the structure of the linear nuclear function is simple, it can be implemented easier. In addition, as the polynomial nuclear function can map the low dimensional feature vectors to the high dimensional feature vectors, it can achieve a better result for a more general nonlinear separable data. Therefore, this paper employs the linear nuclear kernel function, the polynomial nuclear kernel function and the Gaussian kernel function as the kernel functions of the SVR based model. On the other hand, this paper employs the linear nuclear kernel function, the polynomial nuclear kernel function and the Gaussian kernel function as the kernel functions because of having the fair comparisons to the SVR based model.

All the comparisons are under the same simulation conditions. Table 1 shows the values of the $R^2$, the MAE and the MSE as well as the required execution time based on different methods. It can be seen from Table 1 that our proposed method with the Gaussian kernel function yields the highest value of $R^2$ as well as the lowest values of the MAE and the MSE compared to the other methods. Although the required execution time of our proposed method is a little bit higher than that of the kNN based model, the required execution time of our proposed method isstill affordable for the practical applications. Overall, our proposed method is effective for performing the blood glucose estimation.

Table 1
The values of the R2, the MAE and the MSE as well as the required execution time based on different methods.

| Methods | $R^2$ | MAE | MSE | Required execution time (s) |
|---|---|---|---|---|
| Linear regression based model | 0.3377 | 66.8721 | 3799.0142 | 0.031146 |
| kNN based model | 0.2857 | 51.2667 | 4097.4209 | **0.001000** |
| SVR based model with the linear nuclear kernel function | 0.4616 | 43.3352 | 2897.5281 | 0.002000 |
| SVR based model with the polynomial nuclear kernel function | 0.3286 | 49.16248 | 3850.7486 | 0.001001 |
| SVR based model with the Gaussian kernel function | 0.4860 | 42.5893 | 2766.3166 | 0.002000 |
| Decision tree based model | 0.400697 | 47.072093 | 3464.255544 | **0.001000** |
| Random forest based model | 0.405216 | 50.230415 | 3669.034373 | 0.001050 |
| GDBT based model | 0.383300 | 45.802121 | 3564.819113 | 0.117000 |
| BPNN based model | 0.363624 | 51.110924 | 3678.557617 | 1.427433 |
| Our proposed method with the linear nuclear kernel function | 0.552595 | 41.132454 | 2577.136501 | 0.037608 |
| Our proposed method with the polynomial nuclear kernel function | 0.586695 | 37.953387 | 2380.714121 | 0.036821 |
| Our proposed method with the Gaussian kernel function | **0.615004** | **36.887529** | **2217.649565** | 0.037922 |

# 4 Conclusion

This paper proposes a fusion of the multi-regression models based on the histogram information of the blood glucose values in the training set. In particular, the training set is used to establish a preliminary random forest regression model. Then, the blood glucose values in the first validation set can be estimated using this preliminary random forest model. By using the histogram information of the blood glucose values in the training set, the first validation set can be partitioned into 5 subsets. Hence, an individual regression model can be established using the data in each subset of the first validation set. Next, for each feature vector in the second validation set, 5 blood glucose values can be estimated using these 5 random forest models established using the first validation set. By comparing to their reference blood glucose values, the second validation set can be partitioned into 5 subsets. This treats as the labels of the regression models established using the first validation set. As a result, a random forest based classification model is established using the feature vectors in the second validation set and the obtained labels of the regression models established using the first validation set. Finally, by using the random

forest based classification model established using the second validation set, each feature vector in the test set is categorized into one of these 5 classes corresponding to these 5 regression models established using the first validation set. Lastly, the blood glucose value estimated using the corresponding regression model established using the first validation set is employed as the final estimated blood glucose value. Since our proposed method enjoys the advantages of using both the multi-regression models and the classification model for performing the blood glucose estimation, a better regression result is obtained compared to the existing methods under the same simulation conditions.

In future, our proposed method will be applied to estimate other human health indices such as the blood pressure values and the blood lipid values.

# Declarations

Ethical Approval

The ethical approval is not required for this research.

Competing interests

There is no conflict of interest.

Authors' contributions

Yiting Wei is responsible for the methodology, the draft of the paper and implementation of the algorithm.

Weizhi Guo, Yuheng Dai and Qing Liu are responsible for the data acquisition and the verification of the results.

Bingo Wing-Kuen Ling is responsible for the methodology, revising the paper, fund raising and the project management.

Funding

This paper was supported partly by the National Nature Science Foundation of China (no. U1701266, no. 61671163, no. 62071128 and no. 61901123), the Team Project of the Education Ministry of the Guangdong Province (no. 2017KCXTD011), the Guangdong Higher Education Engineering Technology Research Center for Big Data on Manufacturing Knowledge Patent (no. 501130144) and the Hong Kong Innovation and Technology Commission, Enterprise Support Scheme (no. S/E/070/17).

Availability of data and materials

The data is available if it is requested.

# References

1. Sun.H, Saeedi.P, Karuranga.S, et al,“IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045,”2021.
2. Madsbad S, Brock B, Schmitz O. [Postprandial hyperglycemia. Postprandial blood glucose fluctuations, cardiovascular disease and late diabetic complications].[J]. Ugeskrift for Laeger, 2003, 165(33):3149.
3. Mannucci E, Monami M, Pala L, et al. Management of hyperglycemia an type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement from the American Diabetes Association and the European Association for the Study of Diabetes[J]. Diabetologia, 2007, 52(1):17–30.Keck F S, Siegmund T. Continuous Registration of Peritoneal and Subcutaneous Glucose Content by a Combined Microdialyis/Enzymatic Glucose Measuring Device - ScienceDirect[J]. Biosensors '94, 1994:105.
4. SVKR Rajeswari V P. Prediction of Diabetes Mellitus Using Machine Learning Algorithm[J]. Annals of the Romanian Society for Cell Biology, 2021: 5655–5662.Moore, B. The potential use of radio frequency identification devices for active monitoring of blood glucose levels.[J]. Journal of Diabetes Science & Technology, 2009, 3(1):180–183.
5. Huzooree G, Khedo K K, Joonas N. Glucose prediction data analytics for diabetic patients monitoring[C]//2017 1st International Conference on Next Generation Computing Applications (NextComp). IEEE, 2017: 188–195.
6. Wen-Long Q U, Yi-Yi L I, Zhou L. Application of XGBoost algorithm in diabetic blood glucose prediction[J]. Jilin Normal University Journal(Natural Science Edition), 2019.
7. Li M, Fu X, Li D. Diabetes prediction based on XGBoost algorithm[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2020, 768(7): 072093.Jankovic M V, Mosimann S, Bally L, et al. Deep prediction model: The case of online adaptive prediction of subcutaneous glucose[C]// Neural Networks & Applications. IEEE, 2016.
8. Li D, Zhao H, Dou S. A new signal decomposition to estimate breathing rate and heart rate from photoplethysmography signal[J]. Biomedical Signal Processing and Control, 2015, 19:89–95.
9. Cabrera J. On the impact of fusion strategies on classification errors for large ensembles of classifiers[J]. Pattern Recognition, 2006, 39(11):1963–1978.
10. Efron B, Hastie T, Tibshirani J R. Least Angle Regression[J]. Annals of Statistics, 2004, 32(2):407–451.

# Figures

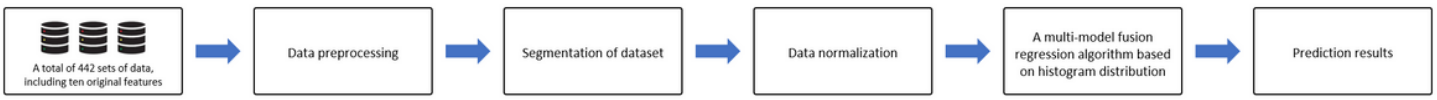



**Figure 1**

The block diagram of our proposed method.

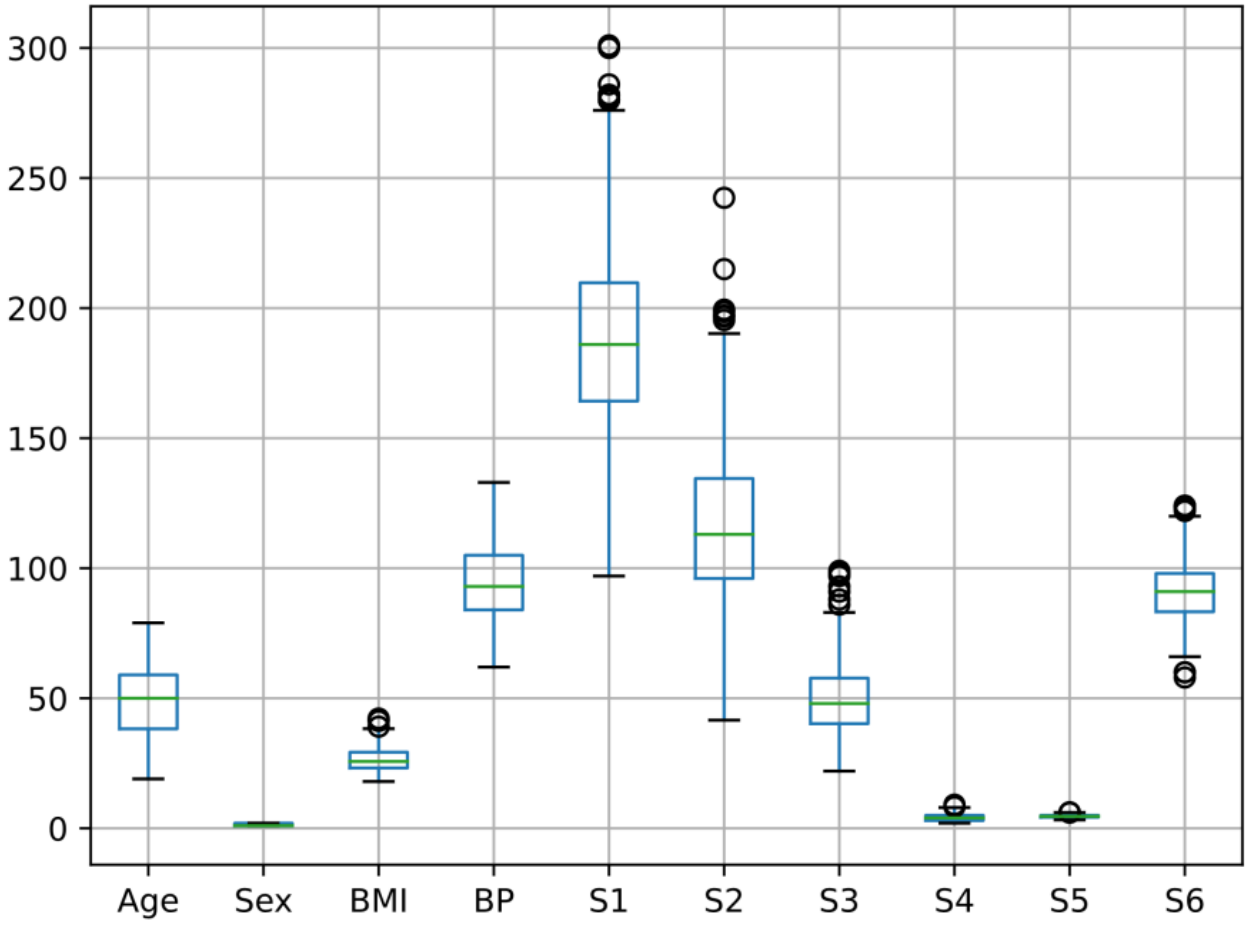
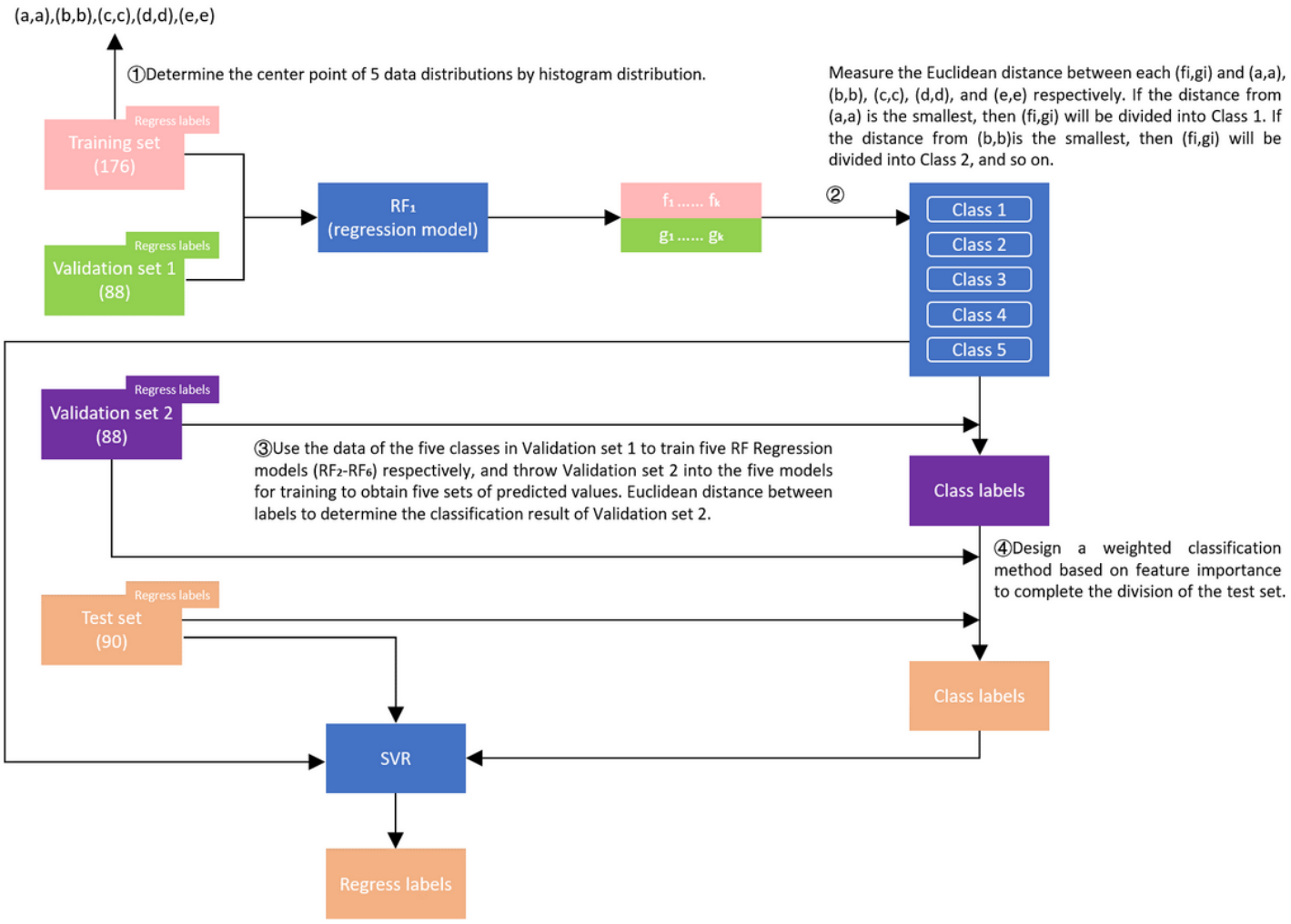



**Figure 2**

Box plot of the features.




## Figure 3

The procedures of our proposed algorithm.

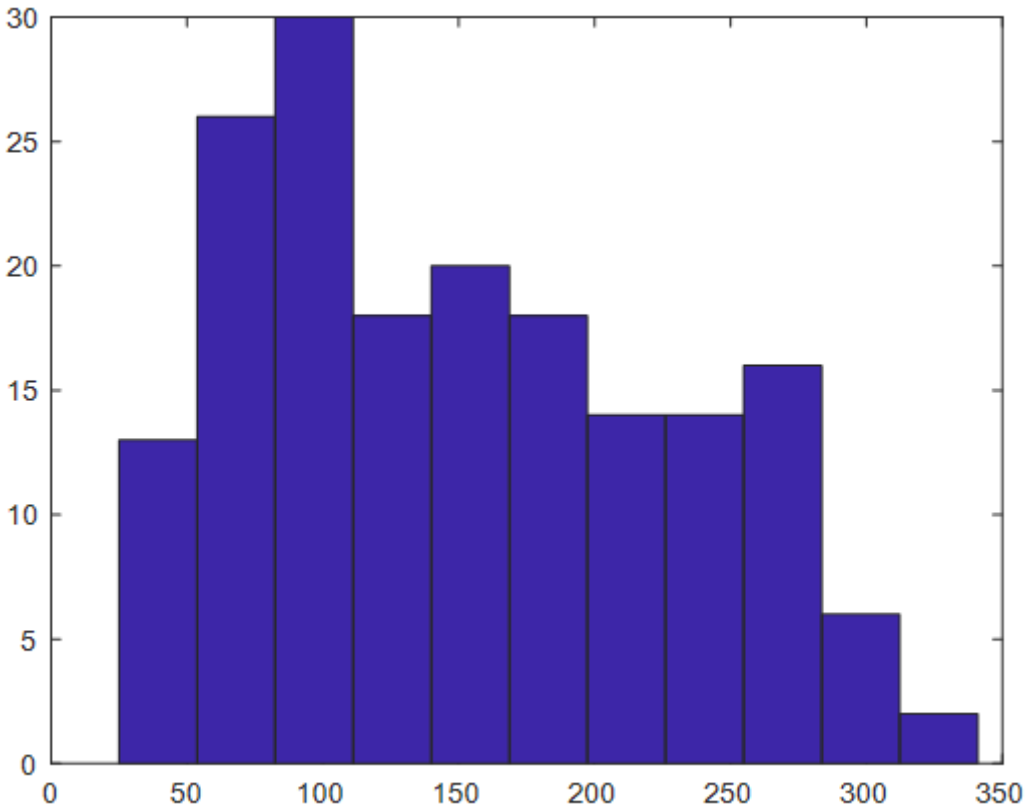



**Figure 4**

The histogram of the blood glucose values in the training set.

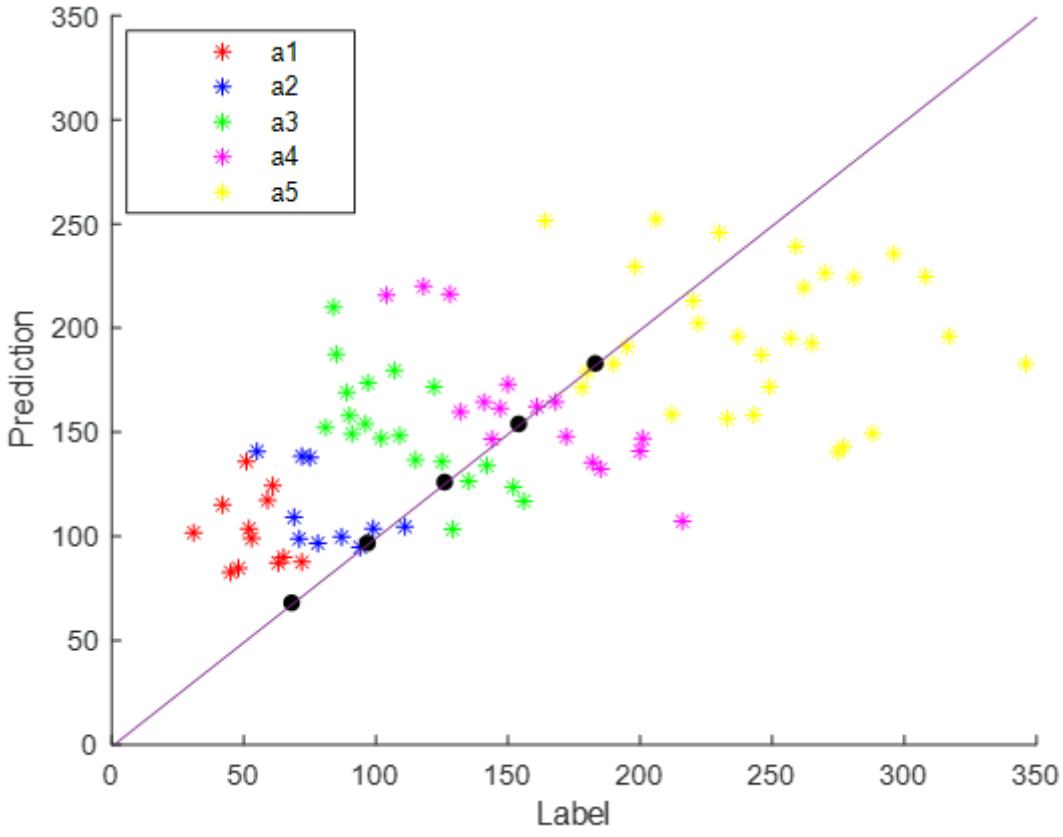



**Figure 5**

The Cartesian coordinates of the pairs of the reference blood glucose values and the estimated blood glucose values in the first validation set