Journal of Universal Computer Science, vol. 24, no. 6 (2018), 797-812
submitted: 5/8/17, accepted: 10/5/18, appeared: 28/6/18 © J.UCS

Integrating Feature Ranking with Ensemble Learning and Logistic Model Trees for the Prediction of Postprandial Blood Glucose Elevation

Jason Chou-Hong Chen
(Gonzaga University, Spokane, WA, USA
chen@gonzaga.edu)

Hsiao-Yen Kang
(Division of Family Medicine, Landseed Hospital, Taiwan
kuanghy@landseed.com.tw
Corresponding author)

Mei-Chin Wang
(Division of Health Management, Landseed Hospital, Taiwan
wangmeria@landseed.com.tw)

Abstract: Postprandial blood glucose (PBG) elevation has been documented as a significant development of diabetes and cardiovascular diseases. Surprisingly, few studies have provided an effective model for predicting PBG elevation. This work presents the classification of PBG in a cohort study via integrating feature ranking with ensemble learning and logistic model trees. We used a cohort dataset that included 1,438 individuals from Landseed Hospital in Taiwan. Data from 2006 to 2013 were collected. To evaluate the performance of the proposed model, four well-known data mining classifiers (Naive Bayes tree algorithm, alternating decision tree, radial basis functions neural network, and Adaboost.M1) were employed in this study. The proposed model provided a reasonably accurate classification for predicting the PBG levels. Twenty-seven risk factors were identified as important risk factors for PBG elevation. The role of PBG should be emphasized and not that of PBG elevation. The predictive factors of PBG must be related to the development of certain diseases.

Keywords: Postprandial Blood Glucose Elevation, Cohort Dataset, Data Mining, Chronic Diseases
Categories: I.2.1, M.4

1 Introduction

Increasing evidence suggests that postprandial blood glucose (PBG) is a contributing factor to the development of cardiovascular diseases and diabetes [Cavalot et al., 2011; Tsujimoto et al., 2016; Kennedy et al., 2017]. A large number of studies have been conducted for predicting/classifying fasting blood glucose and/or PBG elevation [Yamaguchi et al. 2004, Wang and An 2014, Zarkogianni et al. 2015, Wang et al. 2016, Oviedo et al. 2017]. For example, [Tresp et al. 1999] utilized recurrent neural networks and time-series convolution neural networks to predict the blood glucose levels of a patient with diabetes. Their results showed that the recurrent neural network combined with the linear error model can generate the best performance and

outperform both a compartment model and time-series convolution neural-network model. [Wang et al. 2016] used an improved gray GM (1, 1) model to predict PBG in type 2 diabetes with a small amount of data. Their results showed that the improved gray model outperforms the autoregressive (AR) model in predicting blood glucose. [Wang and An 2014] applied the least squares-based AR model for predicting blood glucose levels. Their model can accurately display changes in blood glucose levels to provide an early warning of low blood glucose. [García-Jaramillo et al. 2012] adopted and compared three interval models in the prediction of PBG under uncertainty and intra-patient variability in type 1 diabetes.

However, studies attempting to construct a model for predicting/classifying whether PBG is normal or abnormal are limited. Most existing studies were proposed based on a continuous glucose monitoring system, which is a device that is placed on a patient and used to measure the patient's blood glucose over a specific time period [Oviedo et al. 2017]. The lack of a clear connection between controlled levels of blood glucose and cardiovascular events, diabetes, and all-cause factors has spurred controversial discussions in the field [Wallentin et al., 2016; Tripolt et al., 2016; Völz et al., 2017]. Given that the potential factors are broad categories, we are unable to draw conclusions about the abovementioned relationship. Thus, the influence of relevant risk factors for PBG elevation in a cohort study was determined in this study.

When constructing a model for the classification of PBG, important risk factors must first be discussed and used as predictor variables. Given that numerous risk factors may affect classification accuracy, the selection of key risk factors is crucial in constructing a PBG classification model. Selecting important risk factors has additional benefits. It reduces the number of risk factors appearing in the discovered risk factors, enhances the classification accuracy, and reduces model learning runtime. In the abovementioned studies, few researchers investigated the predictive power of PBG on cardiovascular events and diabetes. We offer insight into whether the predictive factors of PBG play a key role in diabetes and cardiovascular events. Results of this study are expected to help in determining the exact role of blood glucose levels.

In this study, gain ratio attribute evaluation (GR) and information gain attribute evaluation (IG) [Han et al. 2011], two well-known feature ranking methods, were used to rank the importance of features/risk factors. Feature ranking methods can be classified into wrapper and filter method. The wrapper approach uses the method of classification to measure the importance of a set of features; hence, the feature selected depends on the classifier model used. The main disadvantage of the wrapper method is computational complexity and time as each considered feature set must be evaluated with the classifier algorithm used. Conversely, the filter approach is independent of the learning algorithm, computationally simple fast, and scalable. The GR and IG methods belong to filter approaches and have been successfully used in various applications and fields [Saeys et al. 2007].

Most existing studies used only one feature ranking method to rank and select important risk factors. Using only one feature ranking method may not provide stable and effective ranking and selection results. Ensemble learning is a paradigm, in which several intermediate ranking results are generated and combined using ensemble combination rules to finally obtain a single ranking result. It can be used to avoid the unstable ranking/selection results and improve the performance of risk factor

ranking/selection (Dietterich 2000, Polikar 2006, Yang et al. 2010). Borda count, a rank-based scheme, is one of the commonly used voting-based methods of ensemble combination rules [Polikar 2006], and it was used in this study to combine the ranking results of the GR and IG methods.

In this study, we integrated GR and IG feature selection schemes with Borda count ensemble learning and logistic model trees (LMTs) for PBG classification in a cohort study. In the proposed scheme, the GR and IG methods were used to rank the importance of risk factors. Each technique generated a sorted result of risk factors. Borda count then used the sorted results to produce a final ranking of all risk factors. On the basis of the final ranking results, the important risk factors were identified and used as predictors for LMT to construct a final medical diagnosis model for the classification of PBG.LMT is a data mining algorithm for building LMTs, which are classification trees that replace the terminal nodes of a decision tree with logistic regression functions. Each logistic regression is built from all input variables using a stepwise variable selection approach based on the model Akaike information criterion score. This approach gives LMT the theoretical advantage of improved designed splits and enhanced comprehensibility at each node within a tree model [Landwehr et al. 2005; Dancey et al. 2007]. The LMT model has been successfully used for different applications [Landwehr et al. 2005, Dancey et al. 2007, Mahesh et al. 2009, Shoombuatong et al. 2012, Kabir and Zhang 2016, Bui et al. 2016]. However, it has not been used for PBG classification.

The rest of this paper is organized as follows. Section 2 provides a brief introduction about GR, IG, and LMT algorithms. The proposed integrated PBG classification scheme is described in Section 3. Section 4 presents the empirical study results. Finally, the paper is concluded in Section 5.

2 Research Methods

2.1 IG and GR methods

The IG and GR methods evaluate and rank features by calculating the information gain of features that are based on entropy. Entropy is a commonly used measure in information theory, and it characterizes the purity of an arbitrary collection of samples. Entropy is viewed as a measure of a system's unpredictability. The entropy of X is

$$\mathrm{H(X)} = -\sum_{x\epsilon X} p(x) log_2\big(p(x)\big) \tag{1}$$

where $p(x)$is the marginal probability density function for the random variable X.

IG and GR are both decision tree-based feature ranking methods. The GR method is an extension of the IG method. The IG method is defined as the difference in entropy between one parent node and one of its child nodes. Let S be set consisting of *s* data samples with k distinct classes. The information entropy of S is

$$\mathrm{Info(S)} = \sum_{i=1}^{k} p_i log_2(p_i) \tag{2}$$

where p_i is the probability that an arbitrary sample belongs to class k_i and is estimated by s_i/s.

Assume that variable X has d distinct values. Let s_{ij} be the number of samples of class k_i in a subset s_j, which contains those samples in S that have value x_j of X. The expected information entropy based on the partitioning into subsets by X is given by

$$\mathrm{E(X)} = \sum_{i=1}^{k} Info(S) \frac{s_{1i}+s_{2i}+\cdots+s_{ki}}{s} \tag{3}$$

IG is calculated by

$$\mathrm{IG(X)} = \mathrm{Info(S)} - \mathrm{E(X)} \tag{4}$$

The GR method improves the IG method. It evaluates the worth of an attribute by measuring the gain ratio with respect to the class. The difference is that the GR method considers the balance of attribute distribution. In particular, the GR method plays the same role as the IG method, but it provides a normalized measure of a feature's contribution to a classification decision. Thus, GR is less affected by features having a large range of values. The computation for the GR method is one step further to the IG method. It normalizes information gain using a value defined as

$$\mathrm{SpInfo}_X(S) = -\sum_{i=1}^{d} \left(\frac{|S_i|}{|S|}\right) log_2 \left(\frac{|S_i|}{|S|}\right) \tag{5}$$

$\mathrm{SpInfo}_X(S)$ represents the information generated by splitting the training data set S into *d* clusters corresponding to *d* outcomes of a test on the variable/attribute X. GR is defined as

$$\mathrm{GR}(X) = IG(X) - \mathrm{SpInfo}_X(S) \tag{6}$$

2.2 LMT

LMT is a classification model that combines decision tree learning methods and logistic regression. In general, decision trees use a single variable at each tree node to build a model. By contrast, LMT builds a logistic regression model at each node to determine the node's binary split and constructs a piecewise linear approximation of the target function. LMT uses the LogitBoost algorithm [Friedman et al. 2000] for tree construction [Landwehr et al. 2005]. LogitBoost performs forward stage-wise fitting of additive logistic regression models as base learners for the fitting of the logistic models. LMT uses cross-validation to determine a number of LogitBoost iterations to prevent training data overfitting. The LogitBoost algorithm creates an additive model of least-square fits to the given data for each class and uses additive logistic regression of least-squares fits for each class k_i, which has the following form:

$$L_k(x) = \omega_0 + \sum_{i=1}^{n} \omega_i x_i \tag{7}$$

where n is the number of features/factors, and ω_i is the coefficient of the ith component in the observation vector x. The posterior probabilities in the leaf nodes can then be computed by linear logistic regression:

$$\mathrm{p}(k|x) = \frac{exp(L_k(x))}{\sum_{k=1}^{K} exp(L_k(x))} \tag{8}$$

where K is the number if classes, and the least-squares fits $L_k(x)$ are transformed such that $\sum_{k=1}^{K} L_k(x) = 0$.

This hybridized structure reveals that LMT has the advantage of being able to capture the nonlinearities and interaction effects in the dataset for reducing the risk of overfitting. Its flexibility in adapting to the complexity and size of a dataset where the structure of the tree becomes increasingly elaborate is an additional advantage.

3 Proposed Integrated Postprandial Blood Glucose Classification Scheme

This paper integrates GR and IG attribute evaluation methods with Borda count ensemble learning and LMT to propose a PBG classification scheme. The cohort dataset was accessible by the Landseed Hospital in Taiwan. The remaining processes of data cleaning, and data transformation were conducted to improve and obtain enhanced results before data analysis. The flowchart of the proposed scheme is presented in Figure 1. In this figure, the proposed scheme consists of five steps.

The first step of the proposed prediction scheme is to explore the possible risk factors for the prediction of PBG. According to the existing studies related to PBG [Tsujimoto et al., 2016; Kennedy et al., 2017] and suggestions of three clinical physicians, forty-two possible risk factors for the risk of unstable blood glucose were identified (Table 1). The target variable was the type of PBG. Two types of PBG were identified, namely, normal and abnormal PBG of a patient. Thus, the response variable is whether PBG is normal or not.

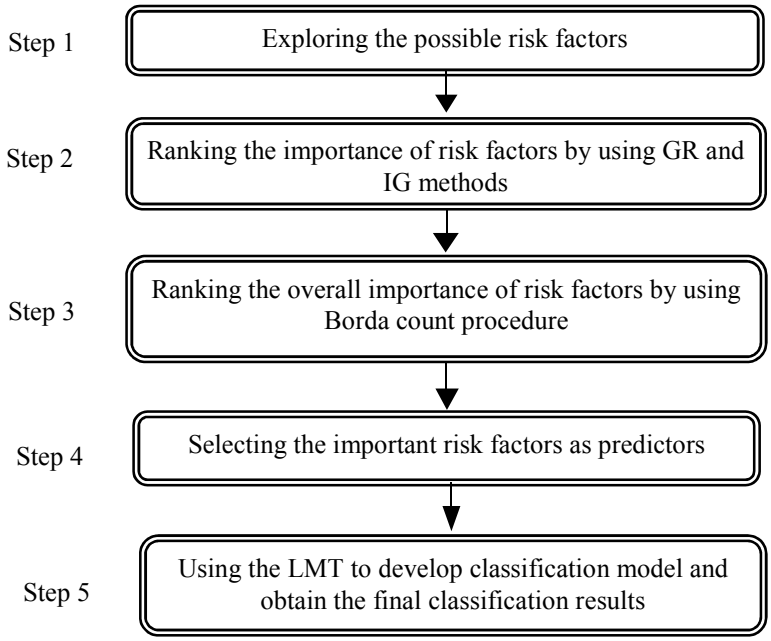

Figure 1: Flowchart of the proposed postprandial blood glucose classification scheme

In the second step, we ranked the importance of predictor variables by using the GR and IG attribute evaluation methods. Tables 2 and 3 show the importance of ranking results of the 42 risk factors by using the GR and IG methods, respectively. GR and IG generated different ranking results. In Table 2, the top five important risk factors using the GR method were urine glucose, casts, anti-HCV, HBsAg, and globulin. Table 3 illustrates that the first five important risk factors using the IG method were urine glucose, HBsAg, casts, HBsAg, and weight. Urine glucose was the most important risk factor in the result of the two methods. This result was also consistent with findings in the literature [Hossain and Park, 2016; Emerging Risk Factors Collaboration 2010; Yamada et al., 2017].

No	Risk factors	No	Risk factors
1	Age	22	LDL, Low-density lipoprotein
2	Gender	23	HBsAg, Hepatitis B surface antigen;
3	WBC, White blood cell	24	Anti-HCV, Hepatitis C antibody;
4	RBC, Red blood cell	25	Urinary protein
5	HGB, Hemoglobin	26	Urine glucose
6	HCT, Hematocrit	27	Ketone body
7	MCV, Mean corpuscular volume	28	Bilirubin
8	MCH, Mean corpuscular hemoglobin	29	Occult blood
9	MCHC, Mean corpuscular-hemoglobin concentration	30	Urine color
10	PLT, Platelet	31	Urine transparency
11	BUN, Blood urea nitrogen)	32	WBCU, White blood cell urine
12	Creatinine	33	RBCU, Red blood cell urine
13	UA, Uric acid	34	Epithelial cell
14	TP, Total protein	35	Urine casts
15	Albumin	36	Urine crystal
16	Globulin	37	Systolic blood pressure
17	Cholesterol	38	Diastolic blood pressure
18	Triglycerides	39	Height
19	GOT, Glutamic-oxalacetic transaminase	40	Weight
20	GPT, Glutamic pyruvic transaminase	41	Waist
21	HDL, High-density lipoprotein	42	Circumference

Table 1: Relevant risk factors for postprandial blood glucose elevation.

No	Risk factors	No	Risk factors
1	Urine Glucose	22	MCH, Mean corpuscular hemoglobin
2	Urine Casts	23	BUN, Blood urea nitrogen)
3	Anti-HCV, Hepatitis C antibody;	24	Epithelial cell
4	HBsAg, Hepatitis B surface antigen;	25	RBCU, Red blood cell urine
5	Globulin	26	HGB, Hemoglobin
6	Weight	27	UA, Uric acid
7	Urinary protein	28	LDL, Low-density lipoprotein
8	TP, Total protein	29	WBC, White blood cell
9	WBCU, White blood cell urine	30	Diastolic blood pressure
10	Triglycerides	31	Urine crystal
11	Ketone body	32	Occult blood
12	Urine transparency	33	HCT, Hematocrit
13	Age	34	Systolic blood pressure
14	RBC, Red blood cell	35	Albumin
15	Creatinine	36	Urine color
16	HDL, High-density lipoprotein	37	MCV, Mean corpuscular volume
17	Bilirubin	38	Gender
18	PLT, Platelet	39	GPT, Glutamic pyruvic transaminase
19	Circumference	40	GOT, Glutamic-oxalacetic transaminase
20	MCHC, Mean corpuscular-hemoglobin concentration	41	Cholesterol
21	Height	42	Waist

Table 2: Importance of ranking results of the 42 risk factors by using the GR method.

No	Risk factors	No	Risk factors
1	Urine glucose	22	UA, Uric acid
2	HBsAg, Hepatitis B surface antigen;	23	LDL, Low-density lipoprotein
3	Anti-HCV, Hepatitis C antibody;	24	Creatinine
4	Urine casts	25	PLT, Platelet
5	Weight	26	HCT, Hematocrit
6	Urinary protein	27	Urine crystal
7	Triglycerides	28	HDL, High-density lipoprotein
8	Globulin	29	WBC, White blood cell
9	Circumference	30	Systolic blood pressure
10	WBCU, White blood cell urine	31	Occult blood
11	Age	32	Bilirubin
12	RBC, Red blood cell	33	Urine color
13	Ketone body	34	RBCU, Red blood cell urine
14	Urine transparency	35	Epithelial cell
15	TP, Total protein	36	MCV, Mean corpuscular volume
16	Height	37	Gender
17	MCH, Mean corpuscular hemoglobin	38	GPT, Glutamic pyruvic transaminase
18	HGB, Hemoglobin	39	Albumin
19	MCHC, Mean corpuscular-hemoglobin concentration	40	Cholesterol
20	BUN, Blood urea nitrogen)	41	GOT, Glutamic-oxalacetic transaminase
21	Diastolic blood pressure	42	Waist

Table 3: Importance of ranking results of the 42 risk factors by using the IG method

The difference in the results of the two feature selection methods, showed that globulin was the fifth risk factor in the GR method but the eighth risk factor in the IG method. From a predicting perspective, the ensemble learning method was used to integrate the results of the two methods.

After ranking the importance of the 42 predictor variables using the GR and IG methods, the third step of the proposed scheme was to use ensemble learning to combine the two ranking results to generate the overall ranking of the importance of the 42 predictor variables. The Borda count procedure was used in this study. For any predictor variable x_k, k $= 1,2, \ldots ,42$, let $B_i(x_k)$ be the order of the predictor variables x_k, which is ranked by the *i*th ranking method R_i in decreasing order. In this study, the Borda count for the predictor variable x_k was B$(x_k) = \sum_{i=1}^{2} B_i(x_k)$. The predictor variable with the largest Borda count was the most important predictor variable. The Borda count value of each predictor variable was used to rank the overall importance of the 42 predictor variables. Table 4 exhibits the overall importance ranking results

of the 42 risk factors by using Borda count ensemble learning. The table shows that Urine glucose, HBsAg, Anti-HCV, Urine casts, and Weight were the first five important risk factors.

Borda count	Risk factors	Rank
2	Urine glucose	1
6	HBsAg, Hepatitis B surface antigen;	2
6	Anti-HCV, Hepatitis C antibody;	3
6	Urine casts	4
11	Weight	5
13	Globulin	6
13	Urinary protein	7
17	Triglycerides	8
19	WBCU, White blood cell urine	9
23	TP, Total protein	10
24	Age	11
24	Ketone body	12
26	RBC, Red blood cell	13
26	Urine transparency	14
28	Circumference	15
37	Height	16
39	MCH, Mean corpuscular hemoglobin	17
39	MCHC, Mean corpuscular-hemoglobin concentration	18
39	Creatinine	19
43	PLT, Platelet	20
43	BUN, Blood urea nitrogen)	21
44	HGB, Hemoglobin	22
44	HDL, High-density lipoprotein	23
49	UA, Uric acid	24
49	Bilirubin	25
51	**LDL, Low-density lipoprotein**	**26**
51	**Diastolic blood pressure**	**27**
58	**WBC, White blood cell**	**28**
58	Urine crystal	29
59	HCT, Hematocrit	30
59	RBCU, Red blood cell urine	31
59	Epithelial cell	32
63	Occult blood	33
64	Systolic blood pressure	34
69	Urine color	35
73	MCV, Mean corpuscular volume	36
74	Albumin	37

Table 4: Overall importance ranking results of the 42risk factors by using Borda count ensemble leaning.

Borda count	Risk factors	Rank
75	Gender	38
77	GPT, Glutamic pyruvic transaminase	39
81	Cholesterol	40
81	GOT, Glutamic-oxalacetic transaminase	41
84	Waist	42

Table 4: (Continued)

In the fourth step, the first q, $q<42$, important predictor variables from the overall ranking result were identified. We considered the overall ranking order of each risk factor and expert suggestions of three clinical physicians after thorough discussion to determine the number of important risk factors/ predictor variables. The Borda count numbers of each risk factor provided useful information. Table 3 demonstrates that the first 27 risk factors were selected as important risk factors/predictor variables. The Borda count numbers of LDL and diastolic blood pressure were the same at 51, and the Borda count number of WBC dramatically decreased to 58. The three clinical physicians also confirmed and suggested that LDL and diastolic blood pressure were more important than WBC in influencing PBG.

In the final fifth step, the identified 27 important risk factors/predictor variables served as the input variables for LMT for PBG classification. The data mining software WEKA, which was developed by [Witten et al. 2016], was utilized to develop the GR and IG attribute evolution methods and LMT models using default settings for each algorithm.

4 Empirical Study

In this study, the PBG dataset of a cohort study provided by Landseed Hospital was used in this study to verify the feasibility and effectiveness of the proposed PBG classification scheme. This cohort study dataset contained 1,438 subject data from 2006 to 2013. Each subject included the records of three physical examinations. The experimental design of this study was to predict the subject with normal PBG in the first two physical examinations who would present abnormal PBG in the third physical examination. In the data, 1,000 subjects exhibited normal PBG in all three physical examinations, and 438 subjects demonstrated abnormal PBG only in the third physical examination. Tenfold cross-validation was used in this study to evaluate the performance of the proposed scheme.

To evaluate the performance of the proposed scheme, four well-known data mining classifiers were employed as competing methods. These methods were Naive Bayes tree (NBtree) algorithm [Kohavi 1996], alternating decision tree [Freund and Mason, 1999], radial basis function neural network [Frank et al. 2014], and Adaboost.M1 method [Freund and Schapire 1996]. These techniques were used to replace the LMT in the proposed method to generate four comparison models. The NBtree, ADtree, Radial basis function neural network (RBFNN), and Adaboost.M1

models were carried out using WEKA software with default settings for each algorithm.

NBtree was formed with the aim to improve the accuracy of the Naive Bayes algorithm, which is a combination of the decision tree algorithm with the Naive Bayes algorithm. At the top of NBtree, it uses the same structure as the structure of a decision tree, whereas the bottom or leaves part uses Naive Bayes classification [Kohavi 1996]. Alternating decision tree (ADtree) provides a special class of decision tree specifically designed for boosting. ADtree maintains the boosting performance and has a smaller and more compact set of rules than NBtree. Instead, of building a forest of decision trees, the boosting procedure of ADtree is incorporated within a single decision tree to facilitate comprehensibility. It consists of alternating layers of decision nodes and prediction nodes starting with a root prediction node [Frank et al. 2014].

RBFNN is an artificial neural network classifier. Its structure is entrenched in clustering, functional approximation, spline interpolation, and mixture models. The input into a RBF network is nonlinear, whereas the output is linear. The output units of a RBF network implement a weighted sum of hidden unit outputs [Frank et al. 2014]. Adaboost.M1 is an extension of the Adaboost method. Adaboost is used as a subroutine to build a classifier with high accuracy in the training set. It applies the classification system repeatedly on the training data; however, on each occasion, it focuses the learning attention on different examples of this set. Once the process has finished, the single classifiers obtained are combined in a final classifier with high accuracy in the training set. Adaboost can only be applied in binary classification problems. Adaboost.M1 is the most simple and natural extension of Adaboost for multiclass classification problems [Freund and Schapire 1996].

To assess the performance metrics of the proposed method, six evaluation measure metrics were used: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F score. The *accuracy index is one of most widely used* performance criterion to measure the percentage of correctly predicted patients. However, sensitivity, specificity, PPV, NPV, and F score are also important measures in medical/healthcare classification issues. The F score combines sensitivity and NPV into a single value and is the harmonic mean of sensitivity and NPV.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{10}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{11}$$

$$\text{PPV} = \frac{TP}{TP+FP} \tag{12}$$

$$\text{NPV} = \frac{TN}{TN+FN} \tag{13}$$

$$\text{Fscore} = \frac{2*Sensitivity*NPV}{Sensitivity+NPV} \tag{14}$$

where TP (true positive) is the number of normal patients correctly classified to the normal patients; TN (true negative) is the total number of abnormal patents correctly classified to abnormal patents; FP (false positive) is the number of normal patients classified to abnormal patients; and FN (false negative) is the abnormal patients classified to normal patients.

The classification results for PBG levels by the proposed scheme and the four competing methods, namely, NBtree, ADtree, RBFNN, and Adaboost.M1, are summarized in Table 5. From the results presented in Table 5, the sensitivity, specificity, accuracy, PPV, NPV, and F score values of the proposed scheme were 86.90%, 63.01%, 79.62%. 84.29%, 67.81%, and 0.8557, respectively. The proposed scheme demonstrated the highest specificity, accuracy, PPV, NPV, and F score values. Although NBtree showed the highest sensitivity of 97.90%, it had the lowest specificity of 9.13% and accuracy of 70.86%. NBtree exhibited better sensitivity values than the proposed method, but it was poor in the other five evaluation measure metrics.

Table 5 reveals that the proposed integrated classification scheme, which integrates feature selection schemes with ensemble learning and LMT, outperformed the four competing models and demonstrated good classification performance. Thus, the proposed scheme was an effective alternative model for the classification of PBG levels.

Models	Sensitivity	Specificity	Accuracy	PPV	NPV	F-Score
Proposed scheme	**86.90%**	**63.01%**	**79.62%**	**84.29%**	**67.81%**	**0.8557**
NBtree	87.50%	45.89%	74.83%	78.69%	61.66%	0.8286
ADtree	85.30%	45.66%	73.23%	78.19%	57.64%	0.8159
RBFNN	83.10%	43.84%	71.14%	77.16%	53.19%	0.8002
Adaboost.M1	97.90%	9.13%	70.86%	71.10%	65.57%	0.8237

Table 5. Classification results of the proposed scheme and the four competing models

5 Conclusion

In this study, we constructed an integrated classification model using GR, IG, Borda count, and LMT methods for PBG classification. With respect to PBG elevation, the suggested five major predictors were urine glucose, HBsAg/ hepatitis B surface antigen, anti-HCV/ hepatitis C antibody, urine casts, and weight. Twenty-seven variables were used as reference risk factors for PBG. Overall, this proposed model provided useful information concerning PBG risk in cohort practice. On the basis of our findings, PBG elevation should be controlled to prevent diabetes and cardiovascular diseases. This study provides insight into the influence of feeding patterns on PBG levels in a large sample, ultimately, it improves cost-effectiveness in health and medicine.

Acknowledgements

This work is supported by the Chung Shan Medical University and LandSeed Hospital: CSMU-LSH106-02.

References

[Bui et al., 2016] Bui, D.T., Tuan, T.A., Klempe, H., Pradhan, B., Revhaug, I.: "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree"; Landslides, 13, 2 (2016), 361-378.

[Dancey et al., 2007] Dancey, D., Bandar, Z.A., McLean, D.: "Logistic model tree extraction from artificial neural networks"; IEEE Transactions on Systems, Man and Cybernetics, 37, 4 (2007), 794–802.

[Dietterich, 2000] Dietterich, T.G.: "Ensemble methods in machine learning"; Lecture Notes in Computer Science, 1857 (2000), 1-15.

[Emerging Risk Factors Collaboration 2010] Emerging Risk Factors Collaboration: "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies"; The Lancet, 375, (2010), 2215-2222.

[Fioravanti et al., 2015] Fioravanti, A., Nikita, K.S.: "Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring"; Medical & Biological Engineering & Computing, 53, 12 (2015), 1333-1343.

[Frank, 2014] Frank, E. Fully supervised training of Gaussian radial basis function networks in WEKA, Department of Computer Science, University of Waikato, (2014).

[Freund and Mason, 1999] Freund, Y., Mason, L.: "The alternating decision tree learning algorithm"; In Proceedings of the Sixteenth International Conference on Machine Learning, Bled, Slovenia, (1999), 124-133.

[Freund and Schapire, 1996] Freund, Y., Schapire, R.E.: "Experiments with a new boosting algorithm"; In Proceedings of the Thirteenth International Conference on Machine Learning, San Francisco, (1996),148-156.

[Friedman et al., 2000] Friedman, J., Hastie, T., Tibshirani, R.: "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)"; The Annals of Statistics, 28, 2 (2000), 337-407.

[García-Jaramillo et al., 2012] García-Jaramillo, M., Calm, R., Bondia, J., Vehí, J.: "Prediction of postprandial blood glucose under uncertainty and intra-patient variability in type 1 diabetes: a comparative study of three interval models"; Computer methods and programs in biomedicine, 108, 1 (2012), 224-233.

[Han et al., 2011] Han, J., Pei, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, MA, (2011).

[Hossain and Park 2016] Hossain, M. F., & Park, J. Y.: "Plain to point network reduced graphene oxide-activated carbon composites decorated with platinum nanoparticles for urine glucose detection"; Scientific reports, 6 (2016), 21009.

[Kabir and Zhang, 2016] Kabir, E., Zhang, Y.: "Epileptic seizure detection from EEG signals using logistic model trees"; Brain Informatics, 3, 2 (2016), 93-100.

[Kennedy et al. 2017] Kennedy, M. W., Fabris, E., Suryapranata, H., & Kedhi, E.: (2017). "Is ischemia the only factor predicting cardiovascular outcomes in all diabetes mellitus patients?"; Cardiovascular diabetology, 16, (1) (2017), 51.

[Kohavi, 1996] Kohavi, R.: "Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree Hybrid"; In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 96 (1996), 202-207.

[Landwehr et al., 2005] Landwehr, N., Hall, M., Frank, E.: "Logistic model trees"; Machine learning, 59, 1-2 (2005), 161-205.

[Mahesh et al., 2009] Mahesh, V., Kandaswamy, A., Vimal, C., Sathish, B.: "ECG arrhythmia classification based on logistic model tree"; Journal of Biomedical Science and Engineering, 2, 6 (2009), 405-411, 2009.

[Oviedo et al., 2017] Oviedo, S., Vehí, J., Calm, R., Armengol, J.: "A review of personalized blood glucose prediction strategies for T1DM patients"; International Journal for Numerical Methods in Biomedical Engineering, 33 (2017), e2833.

[Polikar 2006] Polikar, R.: "Ensemble based systems in decision making"; IEEE Circuits and Systems Magazine, 6, 3 (2006), 21–45, 2006.

[Saeys et al., 2017] Saeys, Y., Inza, I., Larrañaga, P.: "A review of feature selection techniques in bioinformatics"; Bioinformatics, 23, 19 (2017), 2507-2517.

[Shoombuatong et al., 2012] Shoombuatong, W., Hongjaisee, S., Barin, F., Chaijaruwanich, J., Samleerat, T.: "HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees"; Computers in Biology and Medicine, 42, 9 (2012), 885-889.

[Tresp et al., 1999] Tresp, V., Briegel, T., Moody, J.: "Neural-network models for the blood glucose metabolism of a diabetic"; IEEE Transactions on Neural networks, 10, 5 (1999), 1204-1213.

[Tripolt et al. 2016] Tripolt, N. J., Aberer, F., Riedl, R., Hutz, B., Url, J., Dimsity, G., ... & Hafner, F.: "The effects of linagliptin on endothelial function and global arginine bioavailability ratio in coronary artery disease patients with early diabetes: study protocol for a randomized controlled trial"; Trials, 17 (2016), 495.

[Tsujimoto et al. 2016] Tsujimoto, T., Kajio, H., Sugiyama, T.: "Risks for cardiovascular and cardiac deaths in nonobese patients with diabetes and coronary heart disease"; Mayo Clinic Proceedings, 9, (11) 2016; pp. 1545–1554.

[Völz et al. 2017] Völz, S., Svedlund, S., Andersson, B., Li-Ming, G., & Rundqvist, B.: "Coronary flow reserve in patients with resistant hypertension"; Clinical Research in Cardiology, 106 (2017), 151-157.

[Wallentin et al. 2016] Wallentin, L., Lindhagen, L., Ärnström, E., Husted, S., Janzon, M., Johnsen, S. P., ... & Stridsberg, M.: "Early invasive versus non-invasive treatment in patients with non-ST-elevation acute coronary syndrome (FRISC-II): 15 year follow-up of a prospective, randomised, multicentre study"; The Lancet, 388 (2016), 1903-1911.

[Wang and An, 2014] Wang, Y.N., An, B.: "The research least squares based on AR model of glucose prediction"; Advanced Materials Research, 971-973 (2014), 284-287.

[Wang et al., 2016] Wang, Y., Wei, F., Sun, C., Li, Q.: "The research of improved grey GM (1, 1) model to predict the postprandial glucose in Type 2 diabetes"; BioMed Research International, (2016), Article ID 6837052, 6 pages.

[Witten et al., 2016] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: "Data Mining: Practical Machine Learning Tools and Techniques"; Morgan Kaufmann, (2016).

[Yamada et al. 2017] Yamada, Y., Sakuma, J., Takeuchi, I., Yasukochi, Y., Kato, K., Oguri, M., Fujiwara, Y.: "Identification of five genetic variants as novel determinants of type 2 diabetes mellitus in Japanese by exome-wide association studies"; Oncotarget, 8, (46) (2017), 80492.

[Yamaguchi et al., 2006] Yamaguchi, M., Kaseda, C., Yamazaki, K., Kobayashi, M.: "Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining; Medical and Biological Engineering and Computing"; 44, 6 (2006), 451-457.

[Yang et al., 2010] Yang, P., Yang, Y.H., Zhou, B.B., Zomaya, A.Y.: "A review of ensemble methods in bioinformatics"; Current Bioinformatics, 5, 4 (2010), 296-308.

[Zarkogianni et al., 2015] Zarkogianni, K., Mitsis, K., Litsa, E., Arredondo, M.T., Fico, G., Fioravanti, A., Nikita, K.S.: "Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring"; Medical & Biological Engineering & Computing, 53, 12 (2015), 1333-1343.