IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received June 19, 2019, accepted July 14, 2019, date of publication July 19, 2019, date of current version August 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2929866

DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values

QIAN WANG[1], WEIJIA CAO[2], JIAWEI GUO[1], JIADONG REN[1], YONGQIANG CHENG[3], AND DARRYL N. DAVIS[3]

[1]Computer Virtual Technology and System Integration Laboratory of Hebei Province, College of Information Science and Engineering, Yanshan University, Qinhuangdao 066000, China
[2]Qinhuangdao Hospital of Traditional Chinese Medicine, Qinhuangdao 066000, China
[3]Computer Science, University of Hull, Hull HU6 7RX, U.K.

Corresponding author: Qian Wang (wangqianysu@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 61472341, Grant 61772449, Grant 61572420, Grant 61807028, and Grant 61802332.

ABSTRACT As a widely known chronic disease, diabetes mellitus is called a silent killer. It makes the body produce less insulin and causes increased blood sugar, which leads to many complications and affects the normal functioning of various organs, such as eyes, kidneys, and nerves. Although diabetes has attracted high attention in research, due to the existence of missing values and class imbalance in the data, the overall performance of diabetes classification using machine learning is relatively low. In this paper, we propose an effective **P**rediction algorithm for **D**iabetes **M**ellitus classification on **I**mbalanced data with **M**issing values (**DMP_MI**). First, the missing values are compensated by the Naïve Bayes (NB) method for data normalization. Then, an adaptive synthetic sampling method (ADASYN) is adopted to reduce the influence of class imbalance on the prediction performance. Finally, a random forest (RF) classifier is used to generate predictions and evaluated using comprehensive set of evaluation indicators. Experiments performed on Pima Indians diabetes dataset from the University of California at Irvine, Irvine (UCI) Repository, have demonstrated the effectiveness and superiority of our proposed DMP_MI.

INDEX TERMS Diabetes mellitus prediction, machine learning, adaptive synthetic sampling.

I. INTRODUCTION

Pancreas is a most important organ of human body, its produced insulin has an effect on the metabolism of sugar, fat and protein for daily life energy. The blood glucose (i.e. blood sugar) concentration will be high if less or no insulin can be produced and the redundant amount of sugar will be driven out by urine, epitomized as the disease called diabetes mellitus [1]. Diabetes mellitus has become one of the greatest harmful diseases affecting people's life quality as well as creating huge medical costs because of its high incidence rate and complications, such as hypertension, dyslipidemia, stroke, eye disease, kidney disease, etc. [2]. Causes of diabetes are always mysterious even though obesity and lack of exercise play a vital role; it is not only affected by height, weight, hereditary factor but all the factors related to the blood glucose concentration such as diet style. Diabetes is quite common in both developed and developing countries.

In 2007, there were more than 20 million people including adults and children who had suffered diabetes in USA [3]. In India, the diabetic number is expected to increase from more than 30 million in 2000 to about 80 million in 2030. It is also given that, by 2030, 85% of the diabetic patients in the world will be from developing countries [4].

As health care industry develops and generates a mass of useful data such as patient information, electronic medical records, diagnosis and treatment data, and etc., this can serve as a key resource for knowledge extraction that can support decision making and cost reduction. With the continuous development of intelligent analysis methods [5], using intelligence for medical diagnosis has become an unprecedented hot issue [6]. In particular, data mining [7] and machine learning algorithms have gained in strength due to the capability of managing a large amount of data to extract knowledge and

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

make predictions [8]. Researchers have proved that machine learning algorithms [9] such as the classification algorithms of support vector machine, Naïve Bayes, decision tree etc. work in generating better diagnoses for a number of diseases. The ensemble approach, by combining machine learning algorithms, is also proposed to increase the performance and accuracy of diabetes analysis and prediction [10]. Despite the abundantly reported research, diabetes classification remains as a challenging problem for diabetes diagnosis, especially for the early diagnosis and treatment of diabetes which are key to improving the cure of diabetes [11].

This paper focuses on how to achieve a good performance for diabetes classification. Usually, there are missing values and class imbalance problems in medical data, which has a big influence on the classification accuracy. In this work, various factors are comprehensively considered to achieve an improved performance. First, a compensation method is used to improve the data quality and so enable a more effective classification. Then, by applying oversampling technique, data class distribution is balanced, and so avoid problems where the major class unduly biases the classification. The DMP_MI algorithm described here has achieved 87.10% classification accuracy on real diabetes dataset, which outperforms many other algorithm.

The remainder of the paper is organized as follows. Section 2 introduces the related work. Section 3 introduces the foundations of the algorithm. Section 4 develops the DMP_MI algorithm in detail. Section 5 gives results and presents the performance of the DMP_MI algorithm. Section 6 concludes the paper with concluding remarks.

II. RELATED WORK

Various techniques have been applied for diabetes diagnosis and classification. Sahan *et al.* [12] constructed an attribute weighted artificial immune system; the importance of the attributes in class discrimination decided the weights of the attributes, which were then used for Euclidean distance calculation. By using 10-fold cross validation, an accuracy of 75.9% for diabetes classification was obtained. Bozkurt *et al.* [13] compared different methods for diabetes determination using artificial neural network (ANN) and artificial immune system (AIS); they achieved an accuracy of 76.0%. Parashar *et al.* [14] combined linear discriminant analysis (LDA) with support vector machine (SVM) for diabetes diagnosis using the Pima Indians diabetes dataset, and improved the accuracy to 77.6%. Kumari and Chitra [15] also used SVM but combined with radial basis kernel function (RBF) to improve classification accuracy to 78.0%.

Christobel and Sivaprakasam [16] used a Class-wise k Nearest Neighbor (CkNN) method, which interpolated lots of missing values existing in the diabetes dataset through data normalization. The CkNN has a reported accuracy of 78.2%. Khashei *et al.* [17] applied a hybrid classification model of multilayer perceptron, utilizing the soft computing advantages of fuzzy logic, to achieve an accuracy of 80.0%. Farahmandian *et al.* [18] also made some comparisons across techniques such as K-Nearest Neighbor (KNN), Naïve Bayes (NB), Iterative Dichotomiser 3 (ID3), Classification and Regression Tree (CART) and SVM to classify the diabetes data, achieving a best performance of 81.8%. Maniruzzaman *et al.* [19] found that most medical data shows a structure of non-normality, non-linearity and inherent correlation. Thus they adapted a Gaussian process based classification technique with three kernels of linear, polynomial and radial basis, improving the classification accuracy to 82.0%. To the best of our knowledge, Karegowda *et al.* [20] have achieved the best accuracy of 84.7%, through using a hybrid model, which integrated genetic algorithm (GA) and back propagation network (BPN). The following conclusions can be drawn from the above literatures.

(1) The prediction accuracy of diabetes diagnosis remains as a challenging problem and worthy of further research for improvement.

(2) Most of the existing algorithms are based on machine learning methods, indicating that machine learning is effective for diabetes prediction.

(3) Missing value in medical data is a common phenomenon, which has become one of the main problematic factors affecting the classification result.

Typically, the diagnosis of diabetes is regarded as a binary classification problem, i.e. diabetic class and non-diabetic class. The diabetic class is a minor class compared to the non-diabetic population (a major data class). The class imbalance problem in medical data has not been highlighted in the above algorithms. Shigang Liu et al. proposed a fuzzy-based information decomposition (FID) method [21] to simultaneously address the problems of missing values and class imbalance, and viewed these two different problems as a missing data estimation problem. They applied the method on the datasets from different domains such as software engineering, medical related datasets and so on, and obtained better performance than most of the classic algorithms. However, FID recovered the missing values according to the contribution of the observed data and failed to consider the relationship between compensated values and the class label.

Accuracy as an integrity indicator depends on the total number of correct predictions, including diabetic and non-diabetic ones. As an extreme case in imbalance dataset, even if all the samples in the minor class are mispredicted, as long as most samples of the major class are predicted correctly, a high accuracy can still be achieved because of the proportion of the major class [22]. Therefore, we conclude that class imbalance plays a vital role in the classification process [23]–[25], and the accuracy rate alone cannot sufficiently represent classification performance.

To address the above challenge, a prediction algorithm taking into account the problems of missing data and class imbalance has been proposed in the paper. First, the NB method is used to compensate for missing values. NB method is based on learning and can expect to produce better results than simple statistical methods, such as maximum-minimum or mean normalization method. Then, the Adaptive synthetic

sampling method (ADASYN) is adopted to oversample the dataset, increasing the number of the minor class so as to achieve a balance of classes. Finally, a machine learning algorithm Random Forest (RF) is adopted as the classifier, RF draws on the advantage of ensemble learning, namely, several decision trees are integrated to form a new strong predictor [26]. RF adopts voting strategy of the decision trees, and forms an effective way to improve the performance of the classifier.

III. ALGORITHM FOUNATIONS

A. NAÏVE BAYES (NB)

NB is a classification method [27] based on the Bayes theorem with the hypothesis that each feature is independent from every other ones. Let input space $\mathcal{X} \subseteq \mathbf{R}^n$ be the set of n-dimensional samples, and the output space is the set of class labels $\mathcal{Y} = \{c_1, c_2, \ldots, c_K$. The input is the sample $x \in \mathcal{X}$, and the output is the class label $y \in \mathcal{Y}$. X is a random sample defined in the input space $\mathcal{X}$, Y is a random variable defined in the output space Y. $P(X = x \mid Y = c_k)$ is the probability distribution of X given Y, and k $\in \{1, 2 \ldots, K\}$. $P(X = x \mid Y = c_k)$ can be expressed as Formula (1).

$$P(X=x \mid Y=c_k) = P\left(X^{(1)}=x^{(1)}, \cdots, X^{(n)}=x^{(n)} \mid Y=c_k\right) \quad (1)$$

NB method assumes that the features are independent, so it can be obtained as Formula (2).

$$P(X = x|Y = c_k) = \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)}|Y = c_k\right) \quad (2)$$

For a given sample x, according to Bayes theorem [27], the posterior probability $\mathrm{P}(Y = c_k \mid X = x)$ can be calculated through learning as Formula (3). The predicted class label of x is obtained when the posterior probability reaches a maximum value.

$$P(Y=c_k \mid X=x) = \frac{P(X=x \mid Y=c_k) P(Y=c_k)}{\sum_{k=1}^{K} P(X=x \mid Y=c_k) P(Y=c_k)} \quad (3)$$

where $P(Y = c_k)$ is the probability of samples with c_k as its labels. Substitute Formula (2) into Formula (3), and the posterior probability can be further expressed as Formula (4).

$$P(Y = c_k \mid X = x) = \frac{P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(\mathrm{j})} = x^{(j)} \mid Y = c_k\right)}{\sum_{k=1}^{K} P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(\mathrm{j})} = x^{(j)} \mid Y = c_k\right)} \quad (4)$$

Then, NB classifier is calculated as Formula (5), where y is the class label of x, also the value of c_k when the posterior probability reaches the maximum.

$$y = f(x) = arg \max_{c_k} \frac{P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(\mathrm{j})} = x^{(j)} \mid Y = c_k\right)}{\sum_{k=1}^{K} P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(\mathrm{j})} = x^{(j)} \mid Y = c_k\right)} \quad (5)$$

Since the denominator is the same for all labels in Formula (5), the NB classifier can be expressed as Formula (6).

$$y = arg \max_{c_k} P(Y = c_k) \prod_{j=1}^{n} P\left(X^{(j)} = x^{(j)} \mid Y = c_k\right) \quad (6)$$

B. ADAPTIVE SYNTHETIC SAMPLING METHOD (ADASYN)

The ADASYN method is an adaptive data generation method proposed by He *et al.* [28], which can generate samples adaptively to reduce the class imbalance of a dataset. Assume D is a dataset, and there are m samples $\{x_i, y_i\}$ (i $= 1, 2, \cdots,$ m), x_i is a n-dimensional sample, $y_i \in \{0, 1\}$ is a class label, $y_i = 0$ represents the minor class, $y_i = 1$ represents the major class, m_0 represents the number of samples in the minor class, and m_1 represents the number of samples in the major class, where $m_0 \leq m_1, m_0 + m_1 = m$. ADASYN method processes as the following steps.

(1) Evaluate the imbalance degree of all the classes, $d = m_0/m_1, d \in (0, 1)$;

(2) Calculate the total number of samples to be generated, G $= (m_1 - m_0) \times \beta, \beta \in [0, 1]$ represents the expected imbalance degree after data generation. If $\beta = 1$, it means that the samples of the classes are completely balanced after data generation;

(3) For each sample x_i of the minor class, find its k-nearest neighbors in the n-dimensional space. Calculate $\Gamma_i = \Delta_i/k (i = 1, 2, \cdots, m)$, $\Gamma_i \in (0, 1]$.Where Δ_i is the number of the samples which are included in the major class and also the k-nearest neighbors of x_i;

(4) Regularize Γ_i according to $\widehat{\Gamma}_i = \Gamma_i / \sum_i^m \Gamma_i$, then $\widehat{\Gamma}_i$ is a probability distribution, and $\sum \widehat{\Gamma}_i = 1$;

(5) Calculate the sample number of x_i in the minor class to be generated. $g_i = \widehat{\Gamma}_i \times G$;

(6) Select a sample x_j from the k-nearest neighbors of x_i in the minor class. Synthesize new sample S_z, where $S_z = x_i + (x_j - x_i) \times \lambda, \lambda \in [0, 1]$ is a random number.

(7) Repeat Step (6) g_i times to obtain g_i samples of x_i.

C. RANDOM FOREST (RF)

Random forest [29] is an ensemble algorithm based on decision tree (DT) algorithms. Its principle is to generate multiple classification models, where each classifier independently learns and makes predictions. These predictions then collectively vote for final predicted result. The assumption is that better results are expected by using multiple classifiers than individual ones. Commonly used decision trees in random forest ensemble, include the ID3 and CART algorithms. We use CART algorithm in the proposed algorithm, and the corresponding segmentation index is the Gini index. The construction process of random forest is summarized as follows.

(1) Randomly sample with replacement in the original data sets to form n training sets.

(2) For the n training sets, they are used to form n decision tree models. The decision tree will reach its maximum growth without pruning.

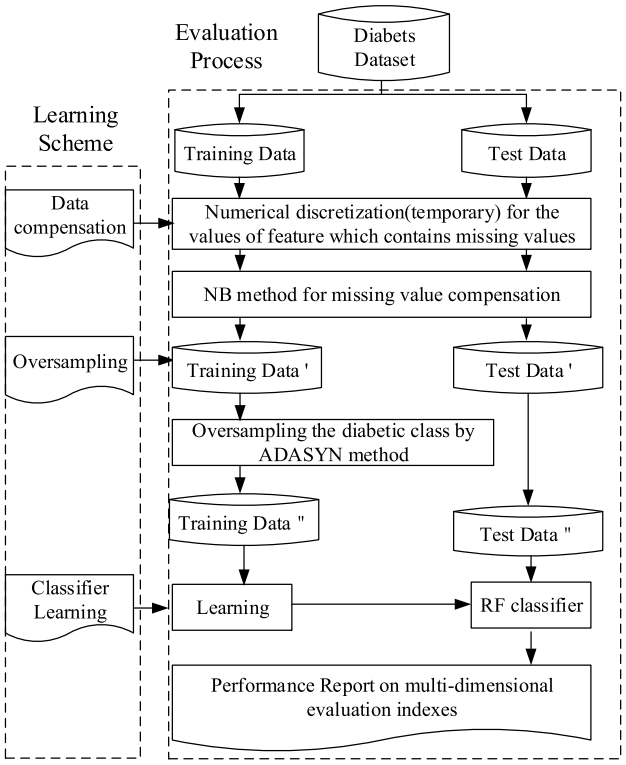


FIGURE 1. The framework of DMP_MI algorithm.

(3) For each individual decision tree, the best feature is selected according to the Gini index for each division.

(4) Repeat (3) until all the training samples of the node belong to the same class.

(5) The n decision trees form a random forest, for the classification problem, the final classification results are determined by the voting of all the n decision trees.

IV. DMP_MI ALGORITHM

The challenges faced by diabetes mellitus prediction can be grouped into three aspects. The first one is the problem of missing data, the second one is the problem of class imbalance, and the third one is the overall performance of the classifier. The proposed DMP_MI algorithm is specially designed as a complete solution to tackle all three challenges. The NB method is used to compensate missing values, which is more targeted than the common statistical methods such as the maximum-minimum or mean normalization method, it trains the samples and compensates for the missing value through prediction, which fully considers the association between the feature being compensated and other features. An oversampling method is adopted to synthesize strongly similar samples through k-nearest neighbors. Comparing to the simple random oversampling by copying the original sample, ADASYN method expands the range of the training samples. Finally, the RF method is adopted to obtain the classification results through the decision tree combination voting mechanism, which has better prediction performance than the single classifier. The framework of DMP_MI algorithm is shown as Figure 1.

TABLE 1. Description of pima indians diabetes database (pidd).

ID	Feature name	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (kg/m^2)
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Age in years
9	Outcome	Class label ('0' or '1')

The steps of DMP_MI algorithm are given below.

(1) As NB is a classification method, to predict the missing values of a continuous-valued feature, numerical discretization should be done first. However, the discretized data is only used for the step of missing value compensation, it is original data without numerical discretization that is used when making the final diabetes classification.

(2) The feature being compensated is viewed as the output label, the remaining features are used as the input. The dataset is divided into test set and training set according to whether the output label is missing or not. If there is an output label, the data will be divided into the training dataset, or, it will be divided into the test dataset. The missing value is compensated according to Formulas (1)-(6). This is repeated until all features are compensated.

(3) According to the ADASYN method, the classes of diabetes only contains diabetic and non-diabetic. ADASYN method is used to oversample the minor class, namely the diabetic class. Then, the numbers of diabetic and non-diabetic samples are consistent with each other, and the purpose of balancing the class data is achieved. The details are described in Section 2.2.

(4) The classification procedure follows the steps described in Section 2.3 to construct n CART trees and form a random forest. The final classification result will be given by the joint voting of multiple trees.

V. EXPERIMENT AND ANALYSIS

A. DATASET AND EXPERIMENTAL ENVIRONMENT

Experiments are performed using the Pima Indians Diabetes Database (PIDD) taken from UCI Repository [30]. Like most medical data, there are missing values and distribution imbalance. The PIDD dataset contains 768 samples, where there are only 392 samples without missing values. Also, diabetic and non-diabetic samples are 268 and 500 respectively, which presents class imbalance. It comprises 8 numeric-valued features and a class label, where the value '0' for a class means negative for diabetes and the value '1' means positive for diabetes. Detailed description of the dataset is shown as Table 1.

Experiments are performed on a PC with Intel(R) Core(TM) i5-4460 at 3.6 GHz CPU and 8GB memory, running on Windows 10. Programs are coded in Python using Pycharm2017 environment on the version of Anaconda3.

TABLE 2. Confusion matrix.

Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

B. PERFORMANCE EVALUATION

The k-fold cross validation approach is adopted as a strategy for preparation of the training and test dataset and to validate the performance. For k-fold cross validation, the experimental data is randomly divided into k subsets. In each experiment, one subset is used as the testing set and the rest k-1 subsets are used as training sets. After a total of k experiments, each subset has been used as testing set for one time. The performance of the model is measured as the average of the results obtained in K experiments.

The classification accuracy is commonly selected as the performance indicator in diabetes analysis using machine learning algorithms. Due to the missing values and class imbalance in the diabetes dataset such as PIDD, accuracy alone is insufficient to evaluate the performance as discussed in the introduction. In order to evaluate and compare our algorithm comprehensively, we carried out the following three scenarios in the experiments,

(1) Conduct classification using RF classifier with and without missing data compensation and oversample technology to determine the effectiveness of the processing for missing data and class imbalance;

(2) Evaluate the results against a set of indicators including accuracy, precision, recall, F1-score and AUC [31] to thoroughly assess the classification performance of DMP_MI algorithm;

(3) Compared with other benchmark algorithms using the readily available accuracy results as the evaluation indicator to verify that DMP_MI algorithm is superior.

The evaluation indicators such as accuracy, precision, recall, F1-score and AUC are defined according to the confusion matrix shown in Table 2. Accuracy indicates the ratio of the number of correctly predicted samples to the total number of samples. Precision indicates the ratio of the number of positive samples correctly predicted to the number of samples predicted as positive ones. Recall indicates the ratio of the number of the positive samples correctly predicted to the number of actual positive samples. F1-score represents the harmonic mean of accuracy and precision. The calculation of accuracy, precision, recall, F1-score and AUC are as below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

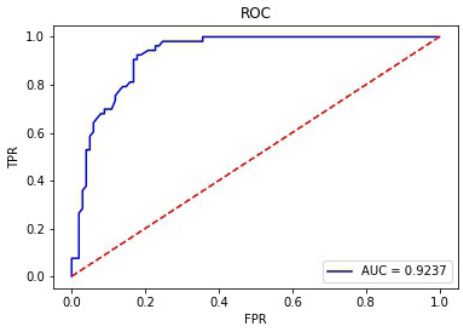


FIGURE 2. The ROC curves and AUC values.

TABLE 3. Comparison between DMP_MI and other algorithms with comprehensive indicators.

k-fold	Algorithm	Accuracy	Precision	Recall	F1-score	AUC
k=10	NB	76.3%	0.759	0.763	0.760	0.819
	SVM	65.1%	0.424	0.651	0.513	0.500
	DT	73.8%	0.735	0.738	0.736	0.751
k=5	RF	78.6%	0.733	0.630	0.673	0.870
	DMP_MI	**87.1%**	**0.806**	0.854	**0.830**	**0.928**
k=10	RF	77.9%	0.688	0.759	0.721	0.841
	DMP_MI	86.2%	0.785	**0.857**	0.816	0.926

AUC (Area Under the Curve) is defined as the area of the ROC (Receiver Operating Characteristic) Curve. The AUC value is the area under ROC curve, which depicts the trade-off between TPR and FPR, that is, by setting varied discrimination thresholds, a series of TPR and FPR is obtained. TPR is the true positive rate, the same as indicator recall, and FPR represents the false positive rate as below.

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

AUC, varying between [0, 1] and generally greater than 0.5, is equivalent to the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative one. If it is equivalent to 0.5, the prediction model is completely ineffective and insignificant for diagnosis application. The larger the value of the indicator, the better the performance of the classifier. The ROC curves and AUC values on PIDD dataset for one time 5-fold cross validation is shown as Figure 2.

Table 3 shows the comparison results between selected traditional machine learning algorithms, RF and the proposed DMP_MI. In RF, there is no compensation of missing data or oversampling of the imbalanced data; the proposed DMP_MI compensates for missing data and solves the class imbalance problem by oversampling the minor class samples. Three benchmark machine learning algorithms, i.e. NB, SVM and DT, were used by Sisodia and Sisodia [32] to perform classification prediction on the diabetes dataset of PIDD with all the indicators defined above. We compare their results with the ones we obtained using the same dataset.

We analyze the experimental results in Table 3 from the following aspects. First of all, for the RF adopted in this paper and the proposed DMP_MI algorithm, the impact of

TABLE 4. Comparison with other algorithms on accuracy.

Reference	Technology	Accuracy
Shigang Liu et al. (2017) [21]	Fuzzy-based Information	67.44%
Sahan et al. (2005) [3]	AWAIS	75.87%
Bozkurt et al. (2014) [4]	ANN,AIS	76.00%
Parashar et al. (2014) [5]	SVM,LDA	77.60%
Kumari et al. (2013) [6]	SVM	78.00%
Christobel et al. (2013) [7]	KNN	78.16%
Khashei et al. (2012) [8]	LDA,QDA,KNN,SVM,ANN,HPM	80.00%
Farahmandian et al. (2015) [9]	SVM,KNN,NB,ID3, CART	81.77%
Maniruzzaman et al. (2017) [10]	LDA,QDA,NB,GPC	81.97%
Karegowda et al. (2011) [11]	GA,BPN	84.71%
Proposed DMP_MI (K=5)	NB-ADAYSN-RF	**87.10%**

the 5-fold or 10-fold cross verification strategies on the effect is not decisive, but the RF is affected more than DMP_MI algorithm, which indicates that the DMP_MI algorithm gives a more reliable performance. Furthermore, by comparing the effect of RF and DMP_MI algorithms, missing value compensation and oversampling play a significant role in the improvement given by DMP_MI. Finally, by comparing the results of DMP_MI algorithm with the algorithms in Reference [32], it can be seen that DMP_MI algorithm achieves the highest accuracy.

Table 4 shows the comparison between the proposed DMP_MI algorithm and a selection of existing algorithms that consider accuracy as the primary performance indicator. It can be observed from the table that DMP_MI outperforms other classifiers in terms of accuracy. It's a common strategy to adopt various benchmark machine learning algorithms collaboratively to combine the advantages of different algorithms in solving the problem of diabetes analysis, but the resulted differences are limited. Among them, References [19] and [20] performed better second and third to the proposed DMP_MI algorithm. The main reason is that, Reference [19] has also considered data structure and solved the problems such as non-normality, non-linearity and inherent correlation, and achieves high-quality data. Reference [20] applied GA and BPN, which are indeed regarded as well performed methods in the field hence the combination of them produces a good effect. However, GA and BPN have high computation time cost on feature selection. Given there are only 8 features in the PIDD dataset, GA and BPN are not suitable solutions. The proposed DMP_MI algorithm takes advantages of this finite limited number of features, and has avoided feature selection process to optimize the efficiency.

VI. CONCLUSION

In this paper, we have proposed a diabetes mellitus classification algorithm targeting the challenging characteristics of missing values and class imbalance problems seen in the diabetes dataset. The algorithm has addressed both the data preprocessing and classification phases. High-quality dataset was obtained by compensating missing values with decreased class imbalance. The validity of the RF classifier is verified by using the k-fold cross validation strategy. Due to the insensitivity of accuracy for imbalanced data, comprehensive indicators for performance evaluation such as accuracy, precision, recall, F1-score and AUC are considered. In the comparison experiment results, the proposed DMP_MI algorithm has outperformed other algorithms on accuracy and other classifier performance indicators, and has shown great potential for diabetes prediction. In the future work, we hope to use the DMP_MI algorithm for the automatic analysis of diabetes with precise severity degree. There are also plans to extend and improve the algorithm to predict and diagnose other type of diseases.

ACKNOWLEDGEMENT

The authors are grateful to valuable comments and suggestions of the reviewers.

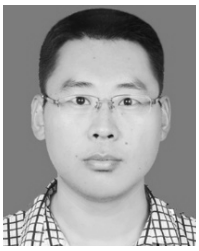
REFERENCES

[1] P. C. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus—A case study," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 1, pp. 8–13, 2015.
[2] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 33, pp. S62–S69, Jan. 2010.
[3] H. N. A. Pham and E. Triantaphyllou, "Prediction of diabetes by employing a new data mining approach which balances fitting and generalization," *Computer and Information Science*, vol. 131. Berlin, Germany: Springer, 2008, pp. 11–26.
[4] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030," *Diabetes Care*, vol. 27, no. 5, pp. 1047–1053, 2004.
[5] X. Wang, D. Bi, and S. Wang, "Fault recognition with labeled multi-category support vector machine," in *Proc. IEEE 3rd Int. Conf. Natural Comput. (ICNC)*, Aug. 2007, pp. 567–571.
[6] B. Zhang, Z. Wei, J. Ren, Y. Cheng, and Z. Zheng, "An empirical study on predicting blood pressure using classification and regression trees," *IEEE Access*, vol. 6, pp. 21758–21768, 2018.
[7] P. S. Kumar and V. Umatejaswi, "Diagnosing diabetes using data mining techniques," *Int. J. Sci. Res. Publications*, vol. 7, pp. 705–709, Jun. 2017.
[8] I. Kavakiotis, O. Tsave, and A. Salifoglou, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, no. 9, pp. 104–116, 2017.
[9] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 9, no. 1, pp. 1–16, 2017.
[10] R. Joshi and M. Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," *Int. Res. J. Eng. Technol.*, vol. 4, no. 10, pp. 426–435, 2017.
[11] S. Ravizza, T. Huschto, A. Adamov, L. Böhm, A. Büsser, F. F. Flöther, R. Hinzmann, H. König, S. M. McAhren, D. H. Robertson, T. Schleyer, B. Schneidinger, and W. Petrich, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nature Med.*, vol. 25, pp. 57–59, Jan. 2019.
[12] S. S. Şahan, K. Polat, S. Güneş, and H. Kodaz, "The medical applications of attribute weighted artificial immune system (AWAIS): Diagnosis of heart and diabetes diseases," in *Proc. Int. Conf. Artif. Immune Syst.* Berlin, Germany: Springer-Verlag, 2005, pp. 456–468.
[13] M. R. Bozkurt, N. Yurtay, Z. Yilmaz, and C. Sertkaya, "Comparison of different methods for determining diabetes," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 22, no. 4, pp. 1044–1055, 2014.
[14] A. Parashar, K. Burse, and K. Rawat, "A comparative approach for Pima Indians diabetes diagnosis using LDA-support vector machine and feed forward neural network," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, no. 11, pp. 378–383, 2014.
[15] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," *Int. J. Eng. Res. Appl.*, vol. 3, no. 2, pp. 1797–1801, 2013.

[16] Y. A. Christobel and P. Sivaprakasam, "A new classwise k nearest neighbor (CKNN) method for the classification of diabetes dataset," *Int. J. Eng. Adv. Technol*, vol. 2, no. 3, pp. 396–400, 2013.
[17] M. Khashei, S. Eftekhari, and J. Parvizian, "Diagnosing diabetes type II using a soft intelligent binary classification model," *Rev. Bioinf. Biometrics*, vol. 1, no. 1, pp. 9–23, 2012.
[18] M. Farahmandian, Y. Lotfi, and I. Maleki, "Data mining algorithms application in diabetes diseases diagnosis: A case study," *MAGNT Res., Tech. Rep.*, vol.3, no.1, pp. 989–997, Jan. 2015.
[19] M. Maniruzzaman, N. Kumar, S. Islam, H. S. Suri, A. S. El-Baz, J. S. Suri, and M. M. Abedin, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed.*, vol. 152, pp. 23–34, Dec. 2017.
[20] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of Pima Indians diabetes," *Int. J. Soft Comput.*, vol. 2, no. 2, pp. 89–96, 2011.
[21] S. Liu, J. Zhang, Y. Xiang, and W. Zhou, "Fuzzy-based information decomposition for incomplete and imbalanced data learning," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1476–1490, Dec. 2017.
[22] Z. Mahmood, D. Bowes, P. C. R. Lane, and T. Hall, "What is the impact of imbalance on software defect prediction performance?" in *Proc. ACM 11th Int. Conf. Predictive Models Data Anal. Softw. Eng.*, Beijing, China, 2015, pp. 1–4.
[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Oct. 2009.
[24] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognit.*, vol. 40, no. 12, pp. 3358–3378, 2007.
[25] T. J. Lakshmi and C. S. R. Prasad, "A study on classifying imbalanced datasets," in *Proc. IEEE 1st Int. Conf. Netw. Soft Comput.*, Aug. 2014, pp. 141–145.
[26] F. Schwenker, "Ensemble methods: Foundations and algorithms [book review]," *IEEE Comput. Intell. Mag.*, vol. 8, no. 1, pp. 77–79, Feb. 2013.
[27] I. Rish, "An empirical study of the Naive Bayes classifier," in *Proc. Workshop Empirical Methods Artif. Intell. (IJCAI)*, vol. 3, no. 22. Seattle, WA, USA: Morgan Kaufmann, 2001, pp. 41–46.
[28] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Hong Kong, Jun. 2008, pp. 1322–1328.
[29] Y. Zhou and G. Qiu, "Random forest for label ranking," *Expert Syst. Appl.*, vol. 112, pp. 99–109, Dec. 2018.
[30] C. L. Blake and C. J. Merz. (1998). UCI Repository of Machine Learning Databases. Department of Information and Computer Science, University of California, Irvine, CA, USA. Accessed: 2003. [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html
[31] H. Tong, B. Liu, and S. Wang, "Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning," *Inf. Softw. Technol.*, vol. 96, pp. 94–111, Apr. 2018.
[32] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.

QIAN WANG received the B.S., M.S., and Ph.D. degrees from the School of Information Science and Engineering, Yanshan University, China, in 2009, 2012, and 2016, respectively. From 2015 to 2016, she was with the University of Hull as a Visiting Scholar.

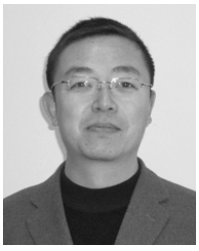
Since 2016, she has been a Lecturer with the School of Information Science and Engineering, Yanshan University, China. Her research interests include data mining, machine learning, software security, and digital health.



WEIJIA CAO received the college engineering degree in computer application and maintenance from Hebei Radio and Television University.

Since 2003, he has been a Staff with the Qinhuangdao Hospital of Traditional Chinese Medicine, where he is currently with the Department of General Services Supervision Section. His research interests include data mining, machine learning, and digital health.

JIAWEI GUO received the B.S. degree in computer science and technology from Shijiazhuang University, China, in 2016. He is currently pursuing the master's degree with the School of Information Science and Engineering, Yanshan University, China. His current research interest includes computer science and technology, especially data mining and machine learning.



JIADONG REN received the B.S. and M.S. degrees from the Northeast Heavy Machinery Institute, in 1989 and 1994, respectively, and the Ph.D. degree from the Harbin Institute of Technology, in 1999.

He is currently a Professor with the School of Information Science and Engineering, Yanshan University, China. His research interests include data mining, complex networks, and software security. He is a Senior Member of the Chinese Computer Society, a member of the IEEE SMC Society, and ACM.

YONGQIANG CHENG received the B.S. and M.S. degrees in control theory and control engineering from Tongji University, Shanghai, China, in 2001 and 2004, respectively, and the Ph.D. degree from the School of Engineering, Design and Technology, University of Bradford, U.K., in 2010

He is currently a Senior Lecturer with the Department of Computer Science, University of Hull, U.K. His current research interests include smart systems and digital health, including ambient living robotics and non-invasive healthcare devices with predictive analysis on the collected data.

DARRYL N. DAVIS is currently a Senior Lecturer and the Director of research with the Department of Computer Science, University of Hull. He is also the Departmental Lead of telehealth, a member of the Intelligent Systems Research Group and the Dependable Systems Research Group. His research interests include telehealth and data mining.