Diabetes Mellitus Prediction Using Multi-objective Genetic Programming and Majority Voting

1st Mehmet Bilgehan Erdem
Dept. of Electrical, Computer, and Software Engineering
University of Ontario Institute of Technology
Oshawa, Canada
bilgehan.erdem@uoit.ca

2nd Zekiye Erdem
Dept. of Electrical, Computer, and Software Engineering
University of Ontario Institute of Technology
Oshawa, Canada
zekiye.erdem@uoit.ca

3rd Shahryar Rahnamayan, SMIEEE
Dept. of Electrical, Computer, and Software Engineering
University of Ontario Institute of Technology
Oshawa, Canada
shahryar.rahnamayan@uoit.ca

***Abstract*— Diabetes is one of the most serious diseases which is becoming increasingly common in recent years. Diabetes can be treated and its consequences are prevented or delayed if predicted timely. This paper investigates an evolutionary computation approach for diabetes prediction. By utilizing the multi-objective Genetic Programming Symbolic Regression, the prediction accuracy level of 79.17% is achieved. Two utilized objectives are namely prediction accuracy and complexity level of the created model (i.e., formula). Moreover, a majority-voting scheme is proposed and compared with other conventional classification algorithms. A widely studied dataset for diabetes prediction, the Pima Indian Diabetes dataset shared in University of California Irvine dataset repository, has been selected for conducting our experimental studies. The work presented here has profound implications for future applications of diabetes prediction and may one help to solve the problem of diabetes by their timely prediction.**

Keywords—Diabetes Prediction, Genetic Programming, SVM, Decision Tree, Linear Discriminant, Logistic Regression, kNN, Majority Voting, Comparative study, Evolution in action.

I. Introduction

Diabetes Mellitus (DM) or commonly referred to as diabetes is a chronic disease caused by inherited and/or acquired deficiency in production of insulin by the pancreas, or by the ineffectiveness of the produced insulin. Such a deficiency results in increased concentrations of glucose in the blood, which in turn can damage many of the body's organs, in particular, the blood vessels and nerves. Especially in early stages, it is often mentioned as "challenging to diagnose" because of its complex inter-dependency to various factors.

World Health Organization (WHO) declares diabetes as a serious and costly disease, which is becoming increasingly common, especially in developing countries and disadvantaged minorities [1]. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation in the world. In 2016, an estimated 1.6 million deaths were directly caused by diabetes globally. Another 2.2 million deaths were attributable to high blood glucose in 2012 [2].

Diabetes can be treated and its consequences avoided or delayed with medication, regular screening, and treatment for complications. For over two decades, a number of researchers had worked on diabetes prediction for early diagnosis and prevention with both various medical and soft computing approaches.

Although the diabetes prediction is widely studied in the relevant literature, to the best of our knowledge, Genetic Programming (GP) Symbolic Regression algorithm was not investigated in this specific area. This paper focuses on the prediction of diabetes using the GP-based Symbolic Regression. A comprehensive comparative study of diabetes prediction with the conventional classification methods, such as, Decision Tree (DT), Linear Discriminant (LD), Quadratic Discriminant (QD), Logistic Regression (LR), Support Vector Machines (SVM), k-Nearest Neighbor (kNN), and finally the proposed majority voting algorithm was conducted and analyzed the results in detail. A widely studied dataset, the Pima Indian Diabetes (PID) dataset, is selected as a benchmark test set for our conduced comparative study [3].

The organization of the paper is as follow, the related studies that used PID dataset for diabetes prediction is reviewed in Section II. Then, detailed information about the aforementioned dataset is given in Section III. The methodology used in this paper, including the comparative study and the majority-voting scheme is explained in Section IV. Afterward, experimental results using PID dataset is presented and discussed in Section V. Concluded remarks, discussion on results and future work are given in the Conclusion Remarks section.

II. Literature Review

Diabetes prediction is a widely studied topic with various datasets. The authors selected one of the most commonly used one, PID dataset from UCI dataset repository [3], to ensure comparison consistency with the relevant literature. The data has been downloaded from Kaggle repository [4].

Kayaer and Yildirim [5] used three different Artificial Neural Network (ANN) structures, namely, MLP, RBF, and GRNN to test the prediction performance of the algorithms on the PID dataset. General Regression Neural Network (GRNN) performed best with an accuracy of 80.21%. In comparison with their related work, ARTMAP IC performed better than their proposed method with an accuracy of 81%. They claimed that

although ARTMAP IC had better classification performance, it also had a more complex NN structure than GRNN.

Priya and Rajalaxmi [6] used a hybrid Feed Forward Back Propagation (FFBP) NN and C4.5 algorithms to predict diabetes on PID dataset. They applied data preprocessing by removing two features with the most missing values, Skin Thickness, and Insulin, namely. They used data mining software, "Rapidminer" and "Weka" for Z score normalization and classification with NN. Although they achieved high accuracy rates with their method, they have not supported any proof for test and validation performances to avoid overfitting.

Rajesh and Sangeetha [7] applied several conventional classification algorithms such as kNN, C-RT, ID3, LDA, C4.5, SVM, and Naïve Bayes on PID dataset. Similar to Priya and Rajalaxmi [6], Rajesh and Sangeetha [7] claims very high accuracy rates (100% with Random Tree) which is not tested with validation dataset. Similar to the previous study, overfitting was not investigated.

Anand, Kirar, and Burse [8] used a High-order Neural Network with Principal Component Analysis (PCA). Similar to the two previous studies, authors claim a very high accuracy rate without validation set, such as around 1^{e-5} MSE. Reported results are highly likely to be over-fitted classification results.

Vijayan and Ravikumar [9] used several different classifications such as Expectation Maximization (EM), kNN, K-means, and ANFIS with adaptive kNN. Both Amalgam kNN and ANFIS adaptive kNN performed best with 80% accuracy. They compared three different types of kNN which they claimed with different k values. However, the k values are not represented in the paper.

Also similar to [9], Radha and Sirinavasan [10] conducted a comparative study using C4.5, SVM, kNN, PNN, and BLR algorithms. They used a data mining software called TANAGRA, which was developed in Delphi programming language. C4.5 classification algorithm performed best with the accuracy of 86% amongst the other classification algorithms in their comparative study. They also conducted a time complexity comparison for the aforementioned algorithms. The fastest algorithm was BLR with 75% accuracy.

Ramesh, Balaji, Iyengar, and Caytiles [11] deployed a deep learning algorithm using Restricted Boltzmann Machine (RBM) as a basic unit, implemented with TensorFlow 1.0 on PID dataset. They used two different diabetes datasets for performance evaluation. Although they achieved 80.99% accuracy with the other dataset, they performed 75% with the PID dataset.

Basha, Balaji, Iyengar, and Caytiles [12] used Linear Discriminant Analysis (LDA), Logistic Regression, Generalized Linear Model (GLMNET), SVM Radial, kNN, Naive Bayes, Regressive Partitioning (rpart), Boosted Tree (C5.0), Bagged CART (treebag), Random Forest (RF), and Generalized Boosted Modeling on PID dataset. They applied preprocessing and dimension reduction on the PID dataset. They also applied Z score normalization.

Table I summarizes a brief description of related studies, which used PID dataset in their diabetes prediction studies.

TABLE I. RELATED RESEARCH WHICH USED PID DATASET

Paper	Employed Algorithms	Result
Kayaer and Yildirim (2003) [5]	MLP, RBF, and GRNN	GRNN 80.21%
Priya and Rajalaxmi (2012) [6]	Hybrid (K-means and C4.5), FFBP NN	FFBP NN 97.93% Overfitting was not investigated
Rajesh and Sangeetha (2012)[13]	kNN, C-RT, ID3, LDA, C4.5, SVM, Naïve Bayes	C4.5 91% Overfitting was not investigated
Anand, Kirar, and Burse (2013) [8]	High-Order NN with PCA	HONN 99.99%
Vijayan and Ravikumar (2014)[9]	ANFIS, kNN, K-means, EM	ANFIS with adaptive kNN 80%
Radha and Sirinavasan (2014)[10]	C4.5, SVM, kNN, PNN, BLR	C4.5 86%
Ramesh, Balaji, Iyengar, and Caytiles (2017) [11]	Deep Learning with RBM	Deep Learning RBM 75%
Basha, Balaji, Iyengar, and Caytiles (2017) [12]	LDA, LR, GLMNET, SVM, kNN, Boosted Tree, Random Forest	LR 78%.

Although the diabetes prediction is widely studied in the relevant literature, to the best of our knowledge, GP algorithm was not investigated in this area. The computational complexity for practical applications has crucial importance in the real-time industrial applications such as IoT. GP has a great potential to reduce the computational complexity significantly in prediction which can be acquired with a single equation, as its prediction model. Furthermore, GP can provide a set of equations with different error rates and complexity levels. Thus, one can choose to predict diabetes by using fewer features. In fact, the utilized multi-objective GP provides the set of Pareto-front solutions with trade-off solutions in terms of prediction accuracy and complexity level of GP formula.

Medical decisions require a high accuracy; for this reason, a majority-voting scheme on different conventional classifiers is employed to improve the prediction accuracy. The details of the conducted experiments are provided accordingly.

III. DATASET INFORMATION

Knowledge about the dataset is important to improve the prediction accuracy. PID dataset is one of the most commonly used datasets for training and testing diabetes prediction models in the relevant literature. The dataset has 8 attributes (i.e., features) and 768 instances. The outcome is logical data which indicates the patient has diabetes or not. All patients are female and older than 21 years old and coming from Pima Indian heritage. Dataset is utilized as it is from the University of California Irvine (UCI) repository [3]. PID dataset features and descriptions are given in Table II.

TABLE II. THE PID DATASET CHARASTERISTICS

No	Name	Description	Type
1	Pregnancy	Number of pregnancy	Integer
2	Glucose	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg)	Integer
3	Blood Pressure	Diastolic blood pressure	Integer
4	Skin Thickness	Triceps skin fold thickness (mm)	Integer
5	Insulin	Two hours serum insulin (muU/ml)	Decimal
6	BMI	Body mass index (weight Kg/height in (mm) ^{2})	Decimal
7	DPF	Diabetes pedigree function	Decimal
8	Age	Age of patient (year)	Integer

Visualization of the data has crucial importance on a better understanding of the data. Fig. 1 shows the histogram charts for the nine attributes of the PID dataset.

Each attribute of PID dataset has different range and boundaries, which can be seen as in the histogram chart given in Fig. 1. The first cell of Fig. 1 shows the number of pregnancies of the Pima Indian women which changes between zero and seventeen pregnancies. The second cell shows the histogram chart of the plasma glucose levels which is distributed between 40-200 mm Hg. Plasma glucose is measured from blood after 2 hours after given patient oral glucose. The third cell is the blood pressure which can be measured manually and varies between 50 to 110 mm Hg. The forth cell is the triceps skin fold thickness measured in mm. There are missing values in the skin thickness feature that can be seen from the total output numbers of this feature. The fifth cell is two-hour serum insulin level measured in muU per ml. The sixth cell is the body mass index which is calculated by dividing weight in kg by the square of the height in meters. The seventh cell is the Diabetes Pedigree Function. The eighth cell is the age in years which is distributed between 21 and 81. The last cell is the outcome of the diabetes functions as 1 diabetes class, and 0 not. In overall, there are 268 diabetes classes and 500 not diabetes classes in the outcome.

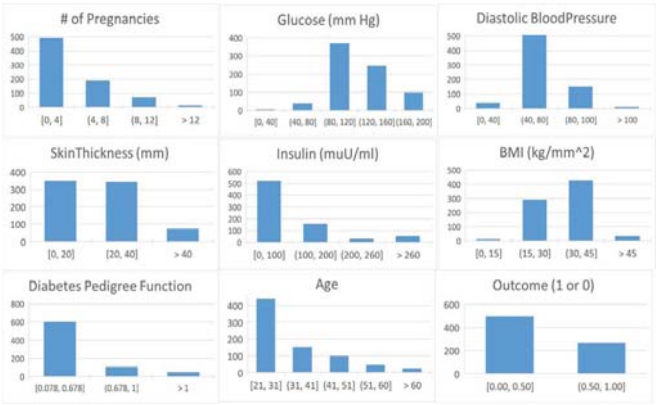

Fig. 1. Histogram charts for the 8 features and the outcome of the PID dataset

IV. METHODOLOGY

This section introduces the classification algorithms used in this study. Firstly, the proposed multi-objective GP Symbolic Regression methodology is covered in sub-section A. Then, the conventional classification algorithms that are employed in this comparative study is given in sub-section B. Finally, the majority-voting scheme used in this paper is explained in the following subsection C.

A. Multi-objective Genetic Programming Symbolic Regression

GP is an evolutionary algorithm to encode computer programs as a set of chromosome those are then evolved using genetic operations such as selection, crossover, and mutation. GP has been successfully used as an automatic programming tool, a machine learning tool and an automatic problem-solving engine [14]. Symbolic regression is a form of regression analysis that searches for a mathematical model that fits best to a given dataset. Symbolic regression via GP is a methodology to automatically generate symbolic models that describe functional relationships on given data [15].

The authors used the Eureqa software [16] from Nutonian Inc. It enabled to have an equation output which user can choose from various mathematical equations (accuracy versus complexity) given as an optimal Pareto-front obtained by the employed multi-objective GP.

B. Comparative Study with Conventional Classification Algorithms

First, the PID dataset is normalized with Z-score normalization based on the descriptive statistics of the dataset. 80% of the data is randomly selected for and the rest 20% is stored for the validation assessments. 5-fold cross validation is applied during training and testing of each competitive algorithm. Then, well-known and widely studied conventional classification algorithms such as Decision Tree (DT), Linear Regression (LR), Quadratic Discrimination (QD), Support Vector Machine (SVM), k Nearest Neighbor (kNN) are employed to classify the PID dataset. The commonly used parameters for the algorithms have been selected based on the relevant literature; e.g. Gini's Diversity Index is used as Split Criterion for the DT, Euclidian Distance metric is used for calculating the neighborhood distances for kNN, k=10, and Linear Kernel is used for SVM. Then, the majority voting ensemble approach is utilized as one of our competitors on the list.

C. Majority Voting Scheme

In order to improve the classification accuracy, the majority-voting scheme of the classifiers, given in the previous section is used. The voting briefly starts after completion of all other comparative classifier algorithms. Then, a counter is set for the majority of each class. The voting stops and the final decision of the voting algorithm are given when the counter reaches the majority threshold. As an example, the majority threshold is set for above 50% of the number of voting classifiers which is set to four out of seven classifiers. Then, when at least four classifiers agree on the same decision, the output of the voting decision is executed.

The formal definition of the voting framework is given in Algorithm 1.

Algorithm 1. Algorithm of proposed Majority Voting

Input: PID Dataset PID(i), M(t) for Methods
Output: Patient Diabetes Decision D(i), Voting Decision V(i), c(i) voting counter, T(Threshold)

```
i=0, V(i)=0, c(i)=0, T=4
  for i=1 to 768
  run M(t) for input PID(i)
      if D(i)=1
        c(i)=c(i)+1
      end if
      if c(i) ≥ T
        V(i)=1
      exit for
      else
        V(i)=0
      end if
  end for
```

V. Results and Discussion

The PID dataset is used for the experimental results. The dataset has 768 instances and 8 features. 80% of the whole dataset was used for training and the rest is reserved for testing. Bi-objective GP is evaluated by using Eureqa software to acquire Symbolic Regression equations. Fig. 12 shows the Pareto-Front solutions of regression equations plotted over two utilized multi-objectives, accuracy versus complexity.

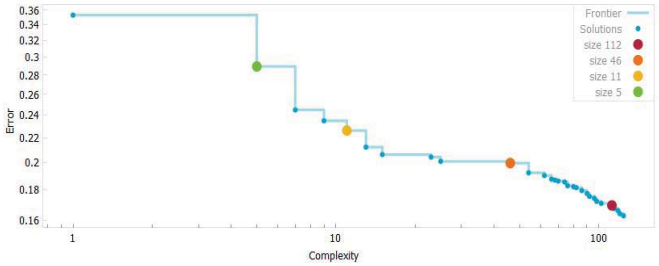

Fig. 2. The resulted Optimal Pareto-front of applied multi-objective GP, error versus complexity of GP formula.

As seen in Fig. 2, GP generates a number of Pareto-front solutions for prediction with a different error rate and complexity. The error range for the generated solutions is between 0.16 and 0.35, while the complexity ranges between 5 and 124. Three equations from the Pareto-front solutions are given for an illustrative example.

The first solution (1) has the Mean Squared Error rate of 0.29 and the minimum complexity of five.

$$O = \text{step}(\text{Glucose}) \tag{1}$$

The second solution (2) has the Mean Squared Error rate of 0.21 and the complexity of 13.

$$O = \text{step}(\text{Pregnancies} + \text{BMI} + \text{DPF} + 1.99*\text{Glucose} - 1.71) \tag{2}$$

The third solution (3) has the Mean Squared Error rate of 0.16 and the complexity of 124.

$$O = \text{step}(\text{BMI} + \text{if}(\text{DPF} - \text{Glucose}, \text{Pregnancies} + \text{DPF}*\cos(\text{BMI}) + \text{if}(\sin(\text{Glucose}), \text{DPF} - a*\sin(\text{Insulin}), \text{Glucose}), 1b + 2*\text{DI} \text{Glucose}^2 - \text{Glucose}*\text{BMI}*\sin(\text{Insulin})) - c) \tag{3}$$

In (3), the coefficients are determined as follows; a= 1.90006018233185, b= 1.48524716076567, c=1.68334822466116. Step function indicates that the output value will be 1 if the output is positive, and 0 otherwise. Step function formula is given in (4) and (5).

$$z = \text{step}(f(x)) \tag{4}$$

$$z = \begin{cases} 1 & \text{if } f(x) \geq 0, \\ 0 & \text{if } f(x) < 0 \end{cases} \tag{5}$$

One of the potential equations from the Pareto-front solutions, (2) is selected due to its balanced structure of both complexity and accuracy for prediction. In other words, Symbolic Regression equation given in (2) can predict diabetes using fewer features with an acceptable, even better accuracy compared to conventional classification algorithms.

A number of models each feature appears in the Optimal Pareto-Front a.k.a. the histogram bar chart of the features which appeared in the Optimal Pareto-Front is shown in Fig. 3. As seen in Fig. 3, Glucose is the feature that is used commonly in most of the generated models.

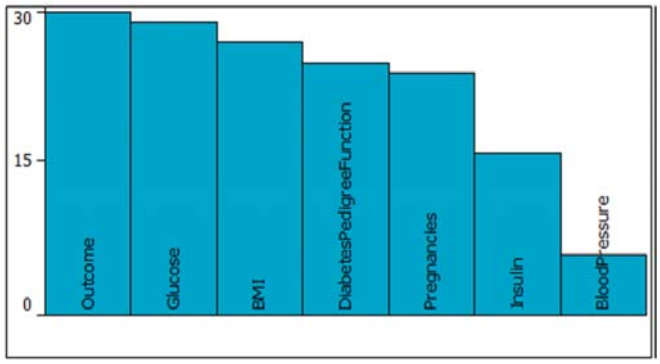

Fig. 3. Number of Models each feature appears in the Optimal Pareto-Front

The number of occurrences of each feature in the Optimal Pareto-Front a.k.a. the frequency bar chart of each feature in the Optimal Pareto-Front is given in Fig. 4. Glucose, DPF, and BMI are the most frequently used features, respectively.

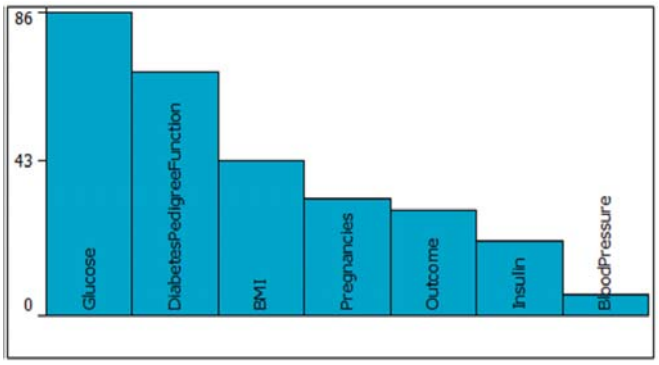

Fig. 4. Number of Occurrences of each feature across all GP models in the Optimal Pareto-Front

Analysis of Fig. 3 and Fig. 4 leads us to interpret the results in a similar way with the innovization approach introduced by Deb [17]. Most important indicators of diabetes can be listed as Glucose, DPF, and BMI which is consistent with the medical literature. Another example of innovization on Pareto-front of the evolutionary multi-objective optimization of GP is (1). This equation can be interpreted as a higher glucose level than average is the strongest indicator of diabetes.

After acquiring GP Symbolic Regression equation given in Eq. 1, the equation is implemented in MATLAB to calculate prediction accuracy. A sample parallel plot for classification accuracy is given in Fig. 5.

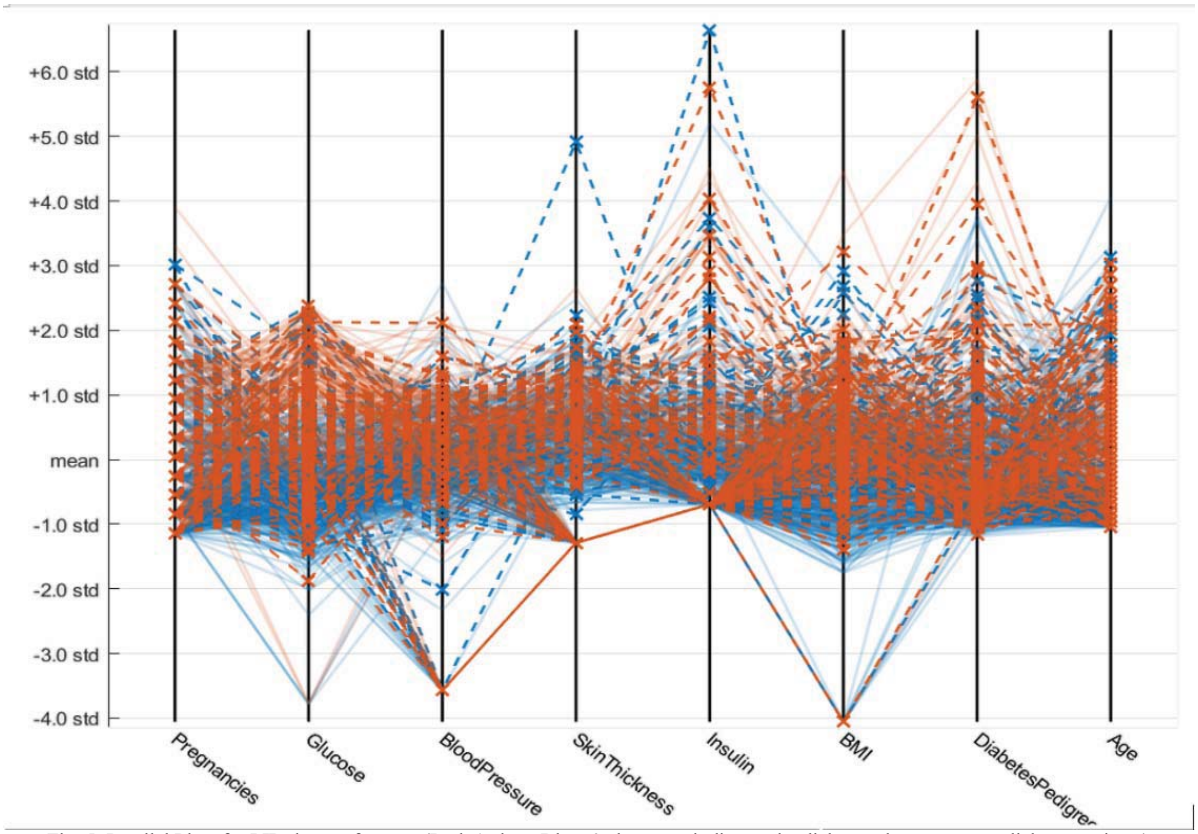

Fig. 5. Parallel Plots for PID dataset features (Red: 1 class, Blue: 0 class, one indicates the diabetes class, zero non-diabetes patients)

As seen in Fig. 5 Insulin and DPF have more outliers than other features.

Accuracy is one of the most common metrics for evaluating classification models. Its verbal definition is the ratio of correct predictions amongst the total number of predictions. Its formal definition is as follows in (4),

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of Predictions}} \quad (4)$$

Fig. 6 shows the results given in Table III for accuracy comparison among selected classification methods, DT, LD, QD, LR, SVM, kNN, GP Symbolic Regression, and Majority-Voting Scheme. Majority-Voting Scheme performs best among the selected classifiers and GP Symbolic Regression performs the second to the best with using only four features.

TABLE III. ACCURACY RESULTS OF COMPARED ALGORITHMS

Methods	**Accuracy**
Quadratic Discriminant	73.43%
K Nearest Neighbor	74.23%
Decision Tree	74.66%
Linear Discriminant	76.81%
Logistic Regression	77.34%
Support Vector Machines	77.69%
Genetic Programming Symbolic Regression	**79.17%**
Majority-Voting Scheme	**81.64%**

a. GP with 4 features

Although Majority-Voting performs better than others do accuracy-wise, it is also the most time and resource-consuming scheme, since it is an ensemble of all the other classification methods used in the study. On the other hand, GP Symbolic Regression is the least costly classifier thanks to its first order linear equation structure. Moreover, GP can be a strong

candidate when the objective is not only the accuracy but also complexity and cost. GP's simple formula can be easily used in embedded systems, especially for real-world applications with limited computational and/or memory capabilities.

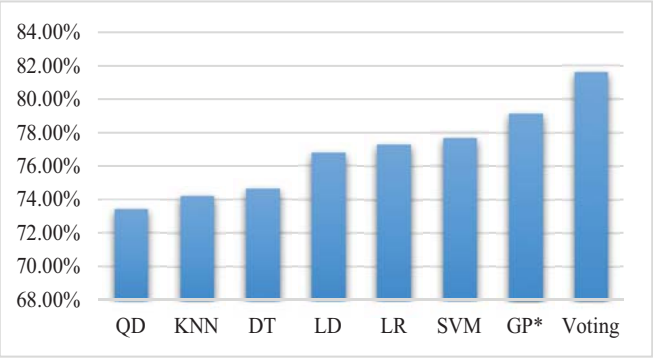

Fig. 6. Accuracy results of utilized classifiers on PID dataset

VI. Conclusion Remarks

Diabetes is a serious disease that is becoming increasingly common. It can be deadly yet preventable if predicted accurately in a timely manner. GP Symbolic Regression is lightweight classifier in terms of computational complexity. Results of the PID dataset verified that GP outperforms other methods even using only 4 out of 8 features.

Although the vast majority of the published work mainly focuses on classification accuracy, one of the overlooked constraints is practical application capabilities of the proposed algorithms. On the contrary, to SVM, kNN, ANN and other heavily computational resource dependent algorithms, GP Symbolic Regression's simple equation classifier can be a potential alternative for prediction with its satisfactory accuracy rate and minimum computational requirements. Therefore, diabetes prediction with GP classifier can find practical applications in the embedded systems due to its potential of lesser Bill of Materials (BoM) costs because of offering the low memory and computational complexities. Moreover, Majority Voting visibly improves the classification accuracy in the cost of higher computation time.

As future work, this comparative study can be extended to cover other healthcare applications also involving a more comprehensive list of machine learning algorithms, including deep learning approaches like CNN and RNN methods.

References

[1] "World Health Organization Fact sheet no.384." URL: https://www.who.int/news-room/fact-sheets/detail/diabetes, p. 5, 2013, access date: 17.01.2019.

[2] "Global Report on Diabetes," Geneva, URL: https://www.who.int/diabetes/global-report/en/, access date: 20.01.2019, 2016.

[3] UCI, "Pima dataset source," 1988. .

[4] "Kaggle Data Base," *https://www.kaggle.com/uciml/pima-indians-diabetes-database*.

[5] K. Kayaer and T. Yildirim, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," *Proc. Int. Conf. Artif. neural networks neural Inf. Process.*, pp. 181–184, 2003.

[6] S. Priya and R. R. Rajalaxmi, "An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network," no. Icon3c, pp. 26–29, 2012.

[7] Rajesh, K. and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," *Int. J. Eng. Innov. Technol.*, vol. Volume 2, no. Issue 3,.

[8] R. Anand, K. Vishnu, and K. Burse, "K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA," *Int. J. Soft Comput. Eng.*, vol. 2, no. 6, pp. 436–438, 2013.

[9] V. Vijayan and A. Ravikumar, "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus," *Int. J. Comput. Appl.*, vol. 95, no. 17, pp. 975–8887, 2014.

[10] P. Radha and B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," vol. 1, no. 6, pp. 334–339, 2014.

[11] S. Ramesh, H. Balaji, and R. D. Caytiles, "Optimal Predictive analytics of Pima Diabetics using Deep Learning," vol. 10, no. 9, pp. 47–62, 2017.

[12] S. Muzamil Basha, H. Balaji, N. S. Ch N Iyengar, and R. D. Caytiles, "A Soft Computing Approach to Provide Recommendation on PIMA Diabetes," *Int. J. Adv. Sci. Technol.*, vol. 106, no. June 2018, pp. 19–32, 2017.

[13] S. V. Rajesh K., "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," *Int. J. Eng. Innov. Technol.*, vol. 2, no. 3, pp. 224–229, 2012.

[14] R. P. W. B. L. N. Freita, *A Field Guide to Genetic Programming - Google Search*, no. March. Published via http://lulu.com and freely available at http://www.gp-field-guide.org.uk, 2008. (With contributions by J. R. Koza)., 2008.

[15] E. J. Vladislavleva, G. F. Smits, and D. den Hertog, "Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming," *IEEE Trans. Evol. Comput.*, vol. 13, no. 2, pp. 333–349, 2009.

[16] "Eureqa (Version 0.98 beta) [Software] Schmidt, M., Lipson, H. (2014) Available from www.nutonian.com," *Eureqa*, 2014. .

[17] A. Gaur and K. Deb, "Adaptive Use of Innovization Principles for a Faster Convergence of Evolutionary Multi-Objective Optimization Algorithms," *Proc. 2016 Genet. Evol. Comput. Conf. Companion - GECCO '16 Companion*, no. July, pp. 75–76, 2016.