Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 157 (2019) 46–54

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12–13 September 2019

Early Detection of Diabetes Mellitus using Feature Selection and Fuzzy Support Vector Machine

Rian Budi Lukmanto[a], Suharjito[a,*], Ariadi Nugroho[a], Habibullah Akbar[a]

[a]*Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

The number of patients that were infected by Diabetes Mellitus (DM) has reached 415 million patients in 2015 and by 2040 this number is expected to increase to approximately 642 million patients. Large amount of medical data of DM patients is available and it provides significant advantage for researchers to fight against DM. The main objective of this research is to leverage F-Score Feature Selection and Fuzzy Support Vector Machine in classifying and detecting DM. Feature selection is used to identify the valuable features in dataset. SVM is then used to train the dataset to generate the fuzzy rules and Fuzzy inference process is finally used to classify the output. The aforementioned methodology is applied to the Pima Indian Diabetes (PID) dataset. The results show a promising accuracy of 89.02% in predicting patients with DM. Additionally, the approach taken provides an optimized count of Fuzzy rules while still maintaining sufficient accuracy.

Keywords: Diabetes Mellitus; Feature Selection; Fuzzy Logic; Support Vector Machine;

1. Introduction

The number of Diabetes Mellitus (DM) patients worldwide has reached 415 million in 2015 [1]. This number is predicted to increase by 55% in 2035 and is claimed to cause one death every 6 seconds [2] [3]. As such DM has become

* Corresponding author. Tel.: +62-812-8400-536.
E-mail address: suharjito@binus.edu

1877-0509
10.1016/j.procs.2019.08.140

an important topic in medical research as to find solutions to fight against the disease, which include employing tools and techniques originated from the field of computer science. Availability of large amounts of labelled medical data related to DM is an advantage for researchers to fight DM. However, employing a traditional method to measure and process large amount of DM data will lead to problems, since the majority of the data have a high level complexity and uncertainty factors [4] [5]. One of the most well-known and powerful method in data processing is the classifier technique. This technique has been used widely to transform traditional approaches to diagnose complex diseases including DM [6].

In this research, we will discuss how Fuzzy SVM as a classifier method will be leveraged to identify and predict DM in a more effective way compared to other classifier methods. To ensure the output of the framework is precise and efficient, F-Score feature selection with pre-processing step is chosen to identify the most valuable features to be analysed in the classification process. Pima Indian Diabetes Dataset, which contains 768 data points of DM patients, is used as the research dataset. Our approach is aligned with other related works which combine F-Score feature selection and Support Vector Machine with promising accuracy result. It has been observed that feature selection is useful in determining the highest SVM classifier accuracy with small number of features [7] [8]. To strengthen the output accuracy, Fuzzy SVM is used to optimize the traditional SVM Classifier. Fuzzy SVM is able to emphasize the support vector node to avoid any redundant training since the crisp sets will be converted to fuzzy sets [9] [10].

The key objective of this experiment is to simplify the classification strategy without sacrificing the accuracy of the output. F-score feature selection is used to identify significant features from Pima Indian Diabetes so that the non-significant features can be removed. Afterwards, Fuzzy SVM is used to optimally classify and train data from the selected features based on feature selection result.

The rest of this paper is organized as follows. Section 2 presents related works that are relevant to this study. In Section 3, the approach in feature selection and Fuzzy SVM technique in classifying DM is explored further. In Section 4, the experiment results and accuracy are evaluated and followed by some concluding remarks in Section 5.

2. Related Works

This section will discuss and share about the past relevant research which focus on classifying DM dataset that computational intelligence concept was chosen as the key method either single or hybrid approaches.

Novel Artificial Bee Colony (ABC) approach has been proposed by Beloufa and Chikh to classify DM dataset [11]. The authors stated that ABC can be an efficient and reliable method to classify diabetes. The authors used Pima Indian Diabetes (PID) datasets and it was concluded that ABC algorithm was a powerful tool for diagnosing Diabetes Mellitus.

Wang et al. proposed an artificial neural network (ANN) compared with multivariate logistics regression (MLR) modelling as a classifier tool for DM [12]. Based on the research, the authors confirmed that computational intelligence methods provides more accurate result compared to regression methods. Decision tree modelling to predict and classify DM has been proposed by Varma et al. [13]. Current traditional decision tree models suffer from a problem of crisp boundaries. The authors proposed to enhance the decision tree with fuzzy computation to prevent the sharp cut-off. Their study used 336 data points, which were tested using the MATLAB tool, and resulted in the accuracy of 75.8%.

Kandhasamy and Balamurali in their study compared each of potential algorithm to find the most optimum classifier technique for DM [14]. The authors compared Decision Tree, K-Nearest Neighbours, Random Forest and Support Vector Machines; and the results show that J48 decision tree generate the lowest accuracy than the other classifier. Zhu et al. proposed a dynamic weighted voting scheme which called multiple factors weighted [6]. In this research, the authors focus on how multiple classifier systems (MCS) can perform in early detection of type 2 Diabetes Mellitus. Meanwhile, the data sets also being used to measure and evaluate the accuracy of the proposed method. In summary they plan to adopt genetic information to establish stronger output for the framework. Lukmanto and Irwansyah employed Fuzzy Hierarchical Model as a generated method of combining Fuzzy System and Analytic Hierarchy Process (AHP) [3]. The main question researched was how Fuzzy + AHP can reduce the number of rules that used to run the inference process and classify the generated output. The authors concluded that their research still need to be improved due to lack in the flow process that impacting the output accuracy.

Re-RX with J48graft which combined with feature selection technique has been proposed by Hayashi and Yukita [15]. They stated that this diagnosis method still facing a complex problem which should be tested on a more recent and complete diabetes datasets to ensure the accuracy. The evaluation of several classifier tools to identify any potential

of DM have been done by Zheng et al. [16]. As the next process of their feature selection process, they also have tried to evaluate several of machine learning framework such as Naïve Bayes, Decision Tree and Support Vector Machine by using sample data of generated electronic health records. Kumar et al. proposed a Support Vector Machine based approach to predict the most discriminatory gene target for Type 2 Diabetes Mellitus [17]. The team has leveraged Support Vector Machine classifier as a feature selector engine to discriminate the gene that has a potential in Type 2 Diabetes Mellitus. The team are confidence that their research giving an insight of how Type 2 Diabetes Mellitus can be predict using Support Vector Machine. Barkana et al. conducted a research related to performance analysis of descriptive statistical features to indicate retinal vessel segmentation caused by Diabetes Mellitus complication called diabetic retinopathy [18]. This research was evaluated using Fuzzy Logic, Artificial Neural Network (ANN) classifier and Support Vector Machine (SVM). All classifiers have been successfully achieved the expected classification accuracies. According to the identified related works from 2013 to 2017 more research needs to be done to prevent the spread of Diabetes Mellitus worldwide by leveraging Computational Intelligence method as the most effective way to detect Diabetes Mellitus.

3. Research Framework

A. Data Pre-processing

The dataset used in this study is the Pima Indian Diabetes (PID) dataset, which was originally came from the National Institute of Diabetes and Digestive and Kidney Diseases (www.niddk.nih.gov). This dataset has been used widely to predict whether a patient has diabetes based on eight diagnostic measurements described below :

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)

The number of missing values in PID dataset is evaluated to determine which features are not reliable to be included in the mining process, and hence is used as a basis to remove incomplete data points. This section describes how this preliminary processing is adopted and implemented as part of the first step in proposed research Framework. Figure 1 describes a systematic process on how each of features from PID are being filtered to assess and list down all missing data. Starting from the pregnancy frequency, once the first feature was successfully being filtered the process move to the second feature and so on. In this computation, we have adopted a basic statistic concept to identify features with large number of missing values, before we move forward with next assessment whether is there any features and feature's instance that potentially can be removed from the mining process

The data pre-processing result need to be assessed to identify which features in PID that produce a lot of number of missing data {Missing}. In this research, we standardize the tolerance percentage for each of the features to be less than or equal to 5%. Number of Times Pregnant was not excluded because zero values indicate no experience of pregnancy rather than a missing value. In the end, out of the eight features we retain six features to be analysed further, removing Triceps Skin Fold Thickness and 2-hour Serum Insulin.

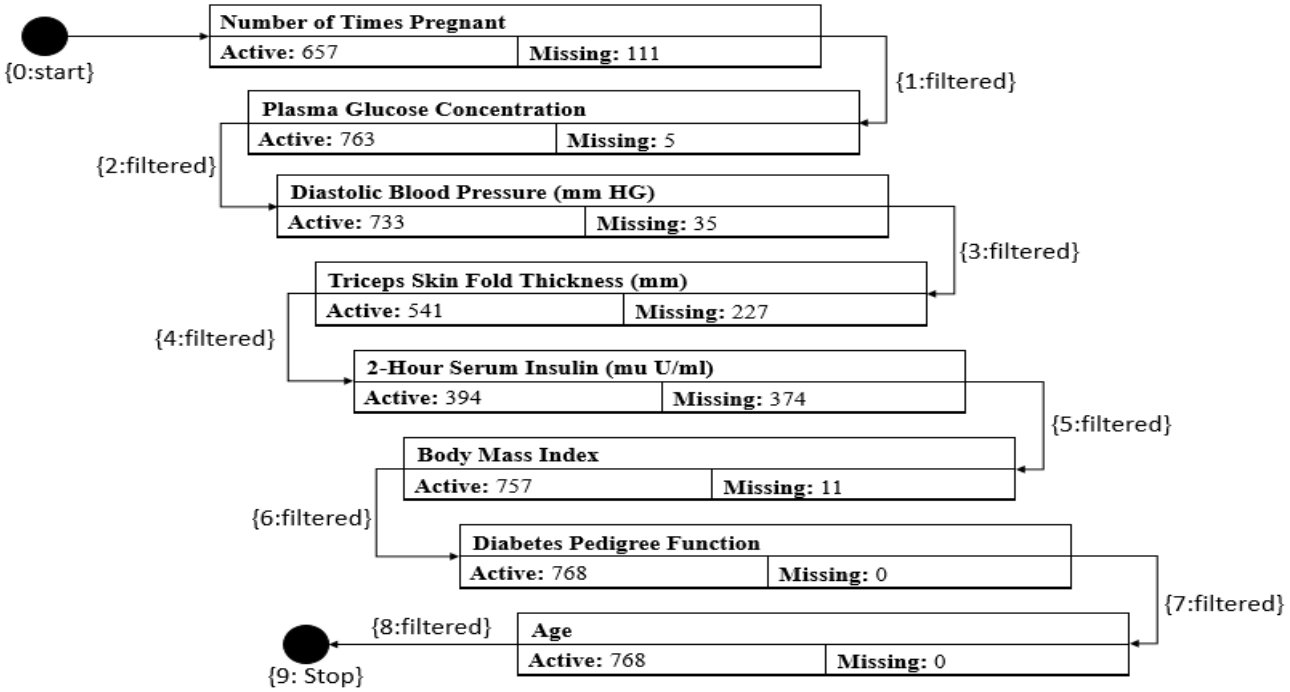


Fig. 1. Data Pre-processing Process

B. Feature Selection

The second step of the proposed Framework is to select highly discriminative features of PID dataset. Once feature's instance with missing values was removed and dataset's features which produce number of missing data in instance with more than 5% being terminated, we move forward to basic and simple technique that measure the distinguishing factor between two classes with real values called F-Score.

The features with relatively high F-scores are considered as informative feature candidates and then are re-studied by the wrapper part that further investigates their contributions in accurate clustering. The features with less F-score are eliminated one at a time from backwards and the performance of the clustering algorithm is observed. In this second step of the proposed framework we are going to get the F-Score values from each of features in PID with following equation :

$$\mathrm{F}(i) \equiv \frac{(x_i^{(+)} - x_i)^2 + (x_i^{(-)} - x_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{+} - x_i^{+})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_+}(x_{k,i}^{-} - x_i^{-})^2} \quad (1)$$

Eq. (1) will be used to get the F-Score value from each of selected features after removing all features and feature's instance with missing data in preliminary data pre-processing step. $x_i^{(+)}, x_i^{(-)}, x_i$ Represent the average value of the total instance in each feature being classified in 3 categories, positive, negative and all data. While, $x_{k,i}^{+}$, $x_{k,i}^{-}$ represent the value in each of feature's instance.

In this research, we leverage Eq. (1) to get the F-Score to discriminative each of PID's features. F-Score is used because it is known to be able to optimize the number of features by removing irrelevant and redundant features [19]. Figure 2 show the exact process of how features selection are implemented in this research.

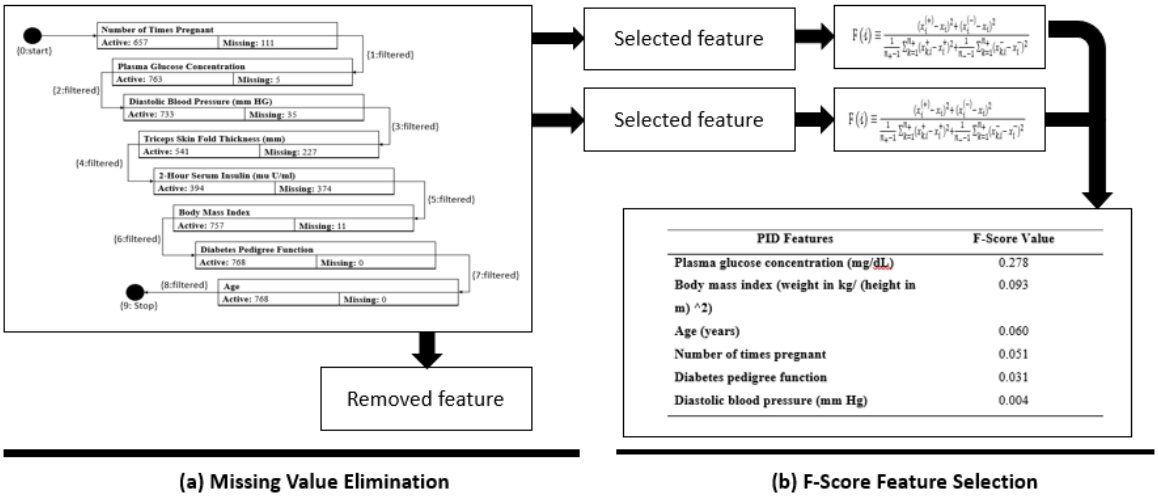


Fig. 2. Proposed Feature Selection Process

C. Classification

Since we have successfully identified the most appropriate features in Pima Indian Diabetes (PID) dataset and clean up all potential noisy data based on pre-processing data and feature selection process, the next step is to start the classification process using Fuzzy SVM techniques. Fuzzy SVM can be used for classification analysis and it is aimed at finding the most optimal hyper plane. In essence, Fuzzy logic is used to classify the level of risks from data, SVM is used to design the fuzzy rules, and the dataset is used to train the SVM using Linear Parameter and test the Fuzzy system.

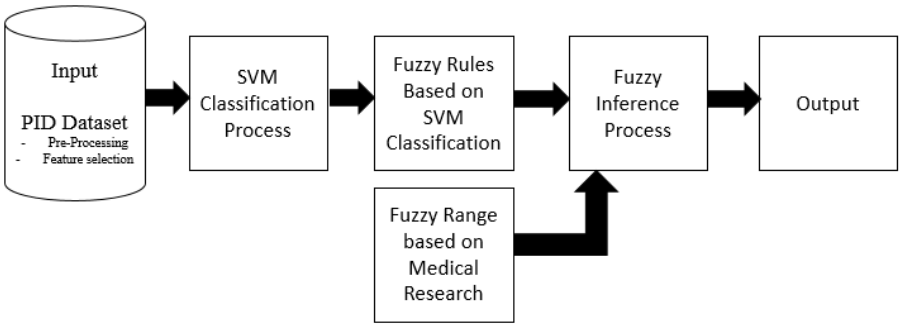
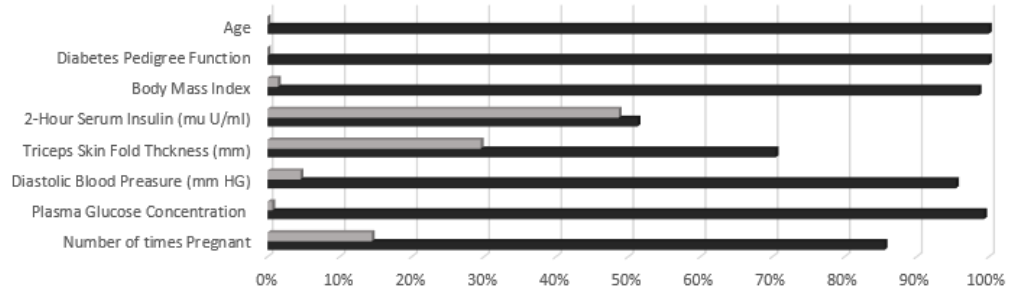


Fig. 3. Proposed Fuzzy SVM Process

All features that were qualified based on the data pre-processing and feature selection are used the in classification process. At this point, Support Vector Machine (SVM) classifier will take the role to classify all data in selected features to generate the fuzzy rules. The standard range for each of the fuzzy set will be based on medical research papers. Figure 3 shows the proposed classification process.

4. Result and Discussion

Based on Figure 4 the number of missing values for the features: 2-hour Serum Insulin, Triceps Skin Fold Thickness and Number of Pregnant are more than 5%, which is {49%; 30%; 14%}, based on criteria of this research tolerance that 3 features need to be removed from the mining process.



	Number of times Pregnant	Plasma Glucose Concentration	Diastolic Blood Preasure (mm HG)	Triceps Skin Fold Thckness (mm)	2-Hour Serum Insulin (mu U/ml)	Body Mass Index	Diabetes Pedigree Function	Age
0 Value	14%	1%	5%	30%	49%	1%	0%	0%
Data Available	86%	99%	95%	70%	51%	99%	100%	100%

Fig. 4. Data Pre-processing Result Assessment

However, Number of Times Pregnant was kept as a feature as discussed previously and we remove 2-Hour Serum Insulin and Triceps Skin Fold Thickness. F-score has been leveraged as a feature selection method to identify and evaluate the level of discriminative for each feature in Pima Indian Diabetes dataset. In this method, given the training vector is $X_k = 1, \dots, m$. With $n_+$and n_- as the positive and negative respective instance. The features with relatively high F-scores are considered as informative feature candidates and then are re-studied by the wrapper part that further investigates their contributions in accurate clustering. The features with less F-score are eliminated one at a time (starting from a lower F-score) and the performance of the clustering algorithm is observed. The generated values of Pima Indian Diabetes dataset is shown in Table 1.

Table 1. F-score values of the selected features in Pima Indian Diabetes dataset

PID Features	F-Score Value
Plasma glucose concentration a 2 hour in an oral glucose tolerance test	0.27
Body mass index (weight in kg/ (height in m) ^2)	0.09
Age (years)	0.06
Number of times pregnant	0.05
Diabetes pedigree function	0.03
Diastolic blood pressure (mm Hg)	0.04

The F-score median value is 0.055 and the mean value is 0.086. In this research, we decide to leverage mean calculation since this approach refers to the standard computation of F-Score in terms of feature selection, which mean there are only 2 features in PID which will be used in classification process. These features are Plasma glucose concentration with 0.278 and Body Mass Index with 0.093 as the defined F-score value.

The inputs to the classifier system are based on pre-processing step and feature selection process. We have reduced the total number of original PID data from 768 to 392. In this process, we split the selected data to 87% (342 data) as a SVM classifier training data to generate the fuzzy rules while 13% (50 data) for data testing to confirm the accuracy of the framework, both data was selected randomly.

Table 2. Table range for Fuzzy Sets

Features	Low	Medium	High
Plasma Glucose Concentration	0 to 130 mg/dL	95 to 230 mg/dL	180 to 350 mg/dL
Body Mass Index	0 to 30 Kg/m2	20 to 40 Kg/m2	30 – 60 Kg/m2

Table 2 shows the standard medical range for both Plasma Glucose Concentration and Body Mass Index based on medical research, while the Fuzzy membership function for each of features have been visualized in Figure 5 and Figure 6.

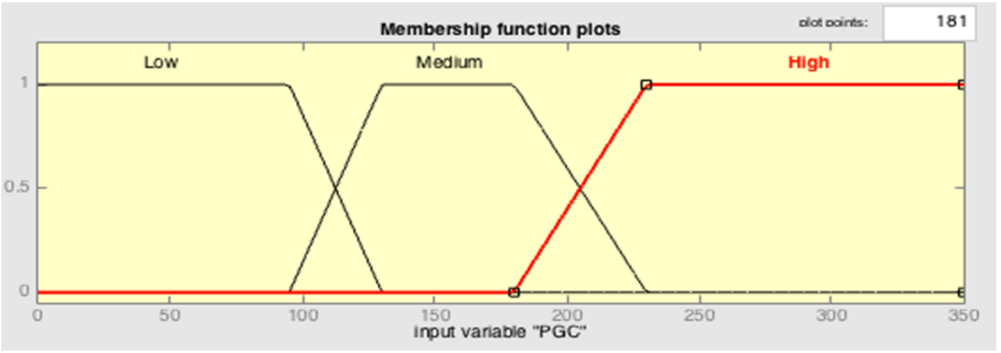


Fig. 5. Fuzzy Membership Function of Plasma Glucose Concentration

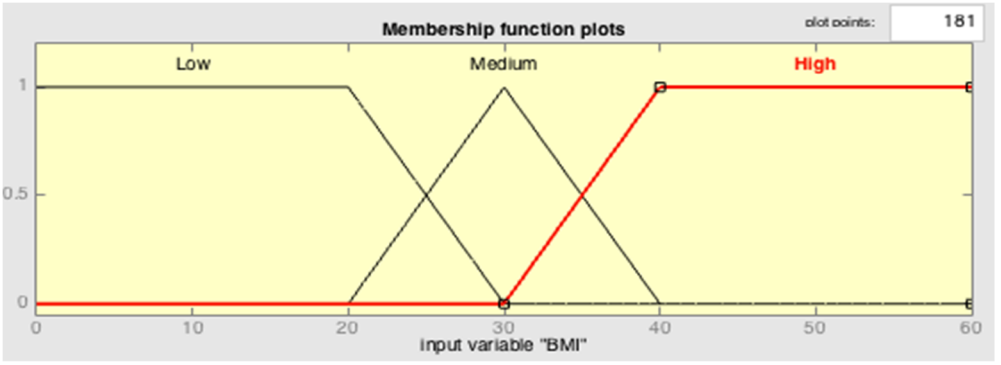


Fig. 6. Fuzzy Membership Function of Body Mass Index

Below is an example of how SVM classifier works in generating the Fuzzy Rules: Based on SVM Classification result the inputs of Plasma Glucose Concentration and Body mass index are 103 and 43.3 then the result of this patients is Negative in Diabetes Mellitus based on the SVM classification result. Hence it can be concluded that IF Plasma Glucose Concentration is Low AND Body Mass Index is high THEN the outcome is Negative in Diabetes.

Table 3. Generated Fuzzy Rules based on SVM Training Data

I/P (IF)	O/P (THEN)
IF PGC is Low AND BMI is Low	8% Positive 92% Negative
IF PGC is Low AND BMI is Medium	36% Positive 64% Negative
IF PGC is Low AND BMI is High	23% Positive 77% Negative
IF PGC is Medium AND BMI is Low	19% Positive 91% Negative
IF PGC is Medium AND BMI is Medium	37% Positive 63% Negative
IF PGC is Medium AND BMI is High	52% Positive 48% Negative
IF PGC is High AND BMI is Low	54% Positive 46% Negative
IF PGC is High AND BMI is Medium	80% Positive 20% Negative
IF PGC is High AND BMI is High	86% Positive 14% Negative

Table 3 shows all generated Fuzzy rules as the result of SVM classification process that will be used as a baseline for Fuzzy inference process. All selected node to create the Fuzzy rules are identified as Support Vector Machine in the classification process. all output with more than 50% of total data will be set as Fuzzy Rules.

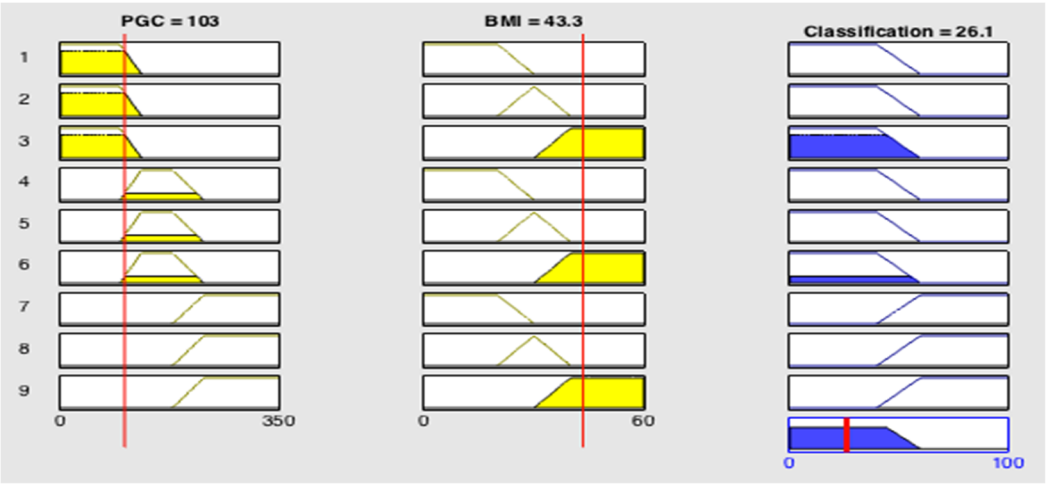


Fig. 7. Sample Rule Viewer

Table 4 shows the testing result of 50 random testing data. In summary, the result of the testing can be classified into: 25 True Positive, 19 True Negative, 1 False Positive and 5 False Negative.

Table 4. Testing Result

Actual Learning Data	**Framework Prediction**	
	Positive	Negative
Positive	25	5
Negative	1	19

This research shows a promising accuracy of 89.02% and has the advantage of optimizing the count of Fuzzy rules while still maintaining sufficient accuracy compared to other research with less accuracy but provide more Fuzzy rules in execution. We also compare this research accuracy result with various classifier techniques in detecting Diabetes Mellitus using the Pima Indian Diabetes dataset reported in other studies, as summarized in Table 5.

Table 5. Related Classifier Method in Diabetes Mellitus.

Method	**Accuracy**
PCA + ANFIS [20]	89%
ANN + FNN [21]	84.24%
Feature Selection + Similarity Classifier [22]	75.97%
Modified Artificial Bee Colony [11]	84.21%
Extreme Machine Learning [23]	77.63%
Hierarchical Fuzzy Classification [24]	79.71%
Re-RX with J48graft [15]	83.83%
Proposed Method	89.02%

The accuracy of this research framework is 89.02% using precision and recall validation technique. Out of 50 data points, 44 data points are correctly classified while 6 data points are wrongly classified. Based on further assessment and investigation of the misclassified data points, five data points are characterized as data points with medium IF PGC AND high BMI or high IF PGC AND low BMI, in which the training result shows non discriminating between negative in Diabetes and positive in Diabetes.

5. Conclusion and Future Works

In this research we have proposed a classification framework to identify and classify the DM dataset using F-

Score Feature Selection and Fuzzy SVM. F-Score feature selection shows is observed to be useful to identify the most valuable features so that it can lead to reducing the number of non-significant features to be further classified. We have also observed that Fuzzy SVM classifier is effective in terms of training the data to generate the Fuzzy rules, so that the proposed Fuzzy Inference can be performed optimally. The experimental result shows a promising result with 89.02% accuracy which is comparable and has the potential to be enhanced in future work. Some of the key opportunities which can be potentially effective to enhance the accuracy of this research is to adopt clustering techniques or employing genetic algorithms as an evolutionary algorithm approach.

References

1. IDF. Diabetes Atlas 7th Edition. ; 2015.
2. Kaul K, Tarr JM, Ahmad SI, Kohner EM, Chibber R. Introduction to diabetes mellitus. In: Diabetes. 2013;: p. 1-11.
3. Lukmanto RB, Irwansyah E. The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. In Procedia Computer Science; 2015; Jakarta: Elsevier. p. 312-319.
4. Sanz JA, Galar M, Jurio A, Brugos A, Pagola M, Bustince H. Medical diagnosis of cardiovascular diseases using an intervalvalued. Applied Soft Computing. 2014;: p. 103-111.
5. Nguyen T, Khosravi A, Creighton D, Nahavandi S. Classification of healthcare data using genetic fuzzy logic system and wavelets. Expert Systems with Applications. 2015;: p. 2184-2197.
6. Zhu J, Xie Q, Zheng K. An improved early detection method of type-2 diabetes mellitus using multiple classifier system. Information Sciences. 2015;: p. 1-14.
7. Ding S. Feature selection based F-score and ACO algorithm in support vector machine. In Knowledge Acquisition and Modeling. 2009;: p. 19-23.
8. Maldonado S, Pérez J, Weber R, & Labbé M. Feature selection for support vector machines via mixed integer linear programming. Information sciences. 2014;: p. 163-175.
9. Abe S. Fuzzy support vector machines for multilabel classification. Pattern Recognition. 2015;: p. 2110-2117.
10. Xie S, Li Z, Hu H. Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization. Gene. 2018;: p. 74-83.
11. Beloufa F, Chikh MA. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer methods and programs in biomedicine. 2013;: p. 92-103.
12. Wang C, Li L, Wang L, Ping Z, Flory MT, Wang G, et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. Diabetes research and clinical practice. 2013;: p. 111-118.
13. Varma KV, Rao AA, Lakshmi TSM, Rao PN. A computational intelligence approach for a better diagnosis of diabetic patients. Computers & Electrical Engineering. 2014;: p. 1758-1765.
14. Kandhasamy JP, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. rocedia Computer Science. 2015;: p. 45-51.
15. Hayashi Y, Yukita S. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked. 2016;: p. 92-104.
16. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. nternational Journal of Medical Informatics. 2017;: p. 120-127.
17. Kumar A, Sharmila DJS, Singh S. SVMRFE based approach for prediction of most discriminatory gene target for type II diabetes. Genomics Data. 2017;: p. 28-37.
18. Barkana BD, Saricicek I, Yildirim B. Performance analysis of descriptive statistical features in retinal vessel segmentation via fuzzy logic, ANN, SVM, and classifier fusion. Knowledge-Based Systems. 2017;: p. 165-176.
19. Ding S. Feature Selection Based F-Score and ACO Algorithm in Support Vector Machine. In Second International Symposium on Knowledge Acquisition and Modeling; 2009; Wuhan: IEEE.
20. Polat K, Güneş S. Artificial immune recognition system with fuzzy resource allocation mechanism classifier, principal component analysis and FFT method based new hybrid automated identification system for classification of EEG signals. Expert Systems with Applications. 2008;: p. 2039-2048.
21. Kahramanli H, Allahverdi N. Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications. 2008;: p. 82-89.
22. Luukka P. Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications. 2011;: p. 4600-4607.
23. Ding S, Zhao H, Zhang Y, Xu X, Nie R. Extreme learning machine: algorithm, theory and applications. Artificial Intelligence Review. 2015;: p. 103-115.
24. Feng TC, Li THS, Kuo PH. Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming. Applied Mathematical Modelling. 2015;: p. 7401-7419.
25. Song Q, Jiang H, Liu J. Feature selection based on FDA and F-score for multi-class classification. Expert Systems with Applications. 2017.