# Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Indians Diabetes dataset


Vaishali R [1 *]
School of Computing Science and Engineering
VIT University
Vellore, India
rvaishali4@gmail.com

Dr. R Sasikala [2]
School of Computing Science and Engineering
VIT University
Vellore, India
sasikala.ra@vit.ac.in

S Ramasubbareddy [3]
School of Computing Science and Engineering
VIT University
Vellore, India
svramasubbareddy1219@gmail.com

S Remya [3]
School of Computing Science and Engineering
VIT University
Vellore, India
remyaakhil@yahoo.com

Sravani Nalluri [3]
School of Computing Science and Engineering
VIT University
Vellore, India
sravani22me@gmail.com



***Abstract*— Diabetes Mellitus is a dreadful disease characterized by increased levels of glucose in the blood, termed as the condition of hyperglycemia. As this disease is prominent among the tropical countries like India, an intense research is being carried out to deliver a machine learning model that could learn from previous patient records in order to deliver smart diagnosis. This research work aims to improve the accuracy of existing diagnostic methods for the prediction of Type 2 Diabetes with machine learning algorithms. The proposed algorithm selects the essential features from the Pima Indians Diabetes Dataset with Goldberg's Genetic algorithm in the pre-processing stage and a Multi Objective Evolutionary Fuzzy Classifier is applied on the dataset. This algorithm works on the principle of maximum classifier rate and minimum rules. As a result of feature selection with GA the number of features is reduced to 4 from 8 and the classifier rate is improved to 83.0435 % with NSGA II in training rate of 70% and 30% testing.**



***Keywords—: Diabetes Mellitus, Genetic algorithm, Multi Objective Evolutionary fuzzy classifier, feature selection***


## I. INTRODUCTION

Diabetes mellitus is a condition of chronic hyperglycemia characterized by the increased levels of glucose in blood due to defects in the secretion of insulin from the Pancreatic Beta cells. The adverse effects of Type 2 diabetes include malfunctioning of organs with permanent damage. The long term effects of diabetes may result in coma, renal failure and retinal failure, pathological destruction of pancreatic beta cells, cardiovascular dysfunction, cerebral vascular dysfunction, peripheral vascular diseases, sexual dysfunction, joint failure, weight loss, ulcer and pathogenic effects on immunity. The reduction in the amounts of insulin causes abnormalities in the levels of carbohydrates and proteins [1]. According to [2] the diabetes mellitus can be characterized by obesity, dysfunction in insulin secretion, abnormal metabolism and increased levels of glucose in excretion. Other symptoms of type 2 diabetes include polyuria, polyphagia, vision malfunction, polydipsia and sudden weight loss. Injection of adequate insulin is the best remedy to treat type 2 Diabetes. There is no long term cure for diabetes but it can be controlled and prevented.


978-1-5090-4642-3/17/$31.00 ©2017 IEEE


On analyzing the facts, [2] has suggested a longitudinal study to understand the cause of insulin dysfunction in the individuals. Availability of standard public datasets has paved possibility for machine learning experts to embark in to the field of diabetes diagnosis exploiting the power of predictive modelling and data analysis.

## II. MOTIVATION

'Dimensionality is a curse to machine learning'. Medical datasets are often larger in dimensions with complex redundant features. The redundancy of features increases the possibility of noise and dependency among the features. The characteristic of a good dataset is to possess less correlated independent variables that are highly correlated to the class or predictive variable. Hence data pre-processing holds a major role in machine learning task with medical datasets. The process of reduction of dimensionality can be carried out either by feature selection or feature extraction. Feature selection concludes a best feature subset out of the existing feature space whereas the feature extraction produces new features extracting essential fragments from the feature space. This work aims to reduce the number of features by a stochastic feature selection method called genetic algorithm. The feature subset selected by the algorithm is evaluated with a classifier and classification accuracy on subset is evaluated with the classification accuracy on original dataset and accuracy achieved in existing literatures. This paper reviews some of the existing literatures on diabetes diagnosis, features present in Pima Indians Diabetes dataset, working of genetic algorithm for feature selection, evaluation of results on multi-objective evolutionary fuzzy classifier and comparison of the results obtained with existing literature.

## III. EXISTING LITERATURE

In this section a collection of literature on the state of art feature selection methods is discussed. In [3] a review on linear dimensionality reduction techniques is made. [4] Has made a stochastic genetic algorithm based approach on the Pima Indians diabetes dataset that has resulted 4/8 features. The features examined with Naïve Bayesian classifier [5] has

produced an accuracy of 78.6957% in (70-30%) supervised learning. [6] Has proposed a hybrid intelligent system for diabetes prediction with GA feature selection and J48 Graft decision tree algorithm. The model produces an accuracy of 74.7826% and ROC of 0.786. [7] Has attempted to diagnose the diabetes with a combination of Genetic algorithm and Multi-Layer Perceptron Neural network. The accuracy of classification in this method is 79.1304%. [8] Has applied Multi objective evolutionary FS algorithm for sales forecasting. [9] Has applied an ant colony based feature selection algorithm in their work. The state of art literature on feature selection shows an opportunity to improve the classification accuracy of the feature selection. Hence we attempt to select features based on Genetic algorithm and test with Multi-Objective evolutionary fuzzy classifier.

## IV. Proposed Work

In this work we combine the Genetic algorithm for feature selection and Multiple Objective evolutionary fuzzy classifier to predict the type 2 diabetes in Pima Indian population. These two algorithms are combined in order to achieve good accuracy in lesser time. The methodology to be followed is discussed and explained with a flowchart. The objectives of the method proposed are as follows

Step 1: Select Pima Indians Diabetes Dataset from the standard data repository of UC Irvine.

Step 2: Genetic feature selection method inspired from [10] is applied to reduce the number of features in the dataset

Step 3: The performance of the Multi-Objective Evolutionary fuzzy classifier is tested on both original dataset and feature reduced dataset.

### A. *Dataset – Pima Indians Diabetes*

The Pima Indians diabetes dataset is a publicly available dataset downloaded from UCI machine learning repository. The dataset comprises of 8 attributes, 768 instances and 1 binary class attribute.

*1) Source of the dataset:* UCI Machine learning repository

*2) About the dataset:* This dataset is a collection of health information from Pima Indian women population of 21 years and above in the region of Arizona and Phoenix

*3) Attributes:* 9

*a)* Number of times pregnant
*b)* Plasma glucose concentration
*c)* Diastolic blood pressure mm Hg
*d)* Triceps skin fold thickness (mm)
*e)* 2-Hour serum insulin (mu U/ml)
*f)* Body mass index $Kgm^{-2}$
*g)* Diabetes pedigree function
*h)* Age in years
*i)* Binary Class variable

### B. *Flowchart chart of the proposed work*

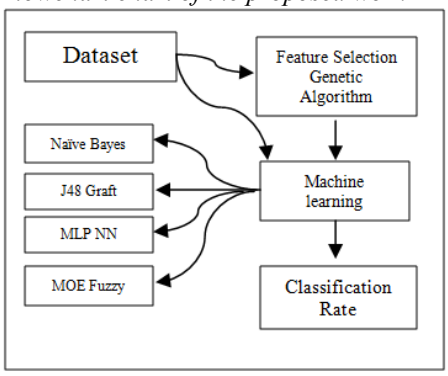


Fig.1. Flowchart of the proposed feature selection and classification

## V. Genetic Algorithms for Feature Selection

This paper implements the genetic algorithm described by [10] for feature selection. As exhaustive search methods are expensive on large feature sets, we choose a stochastic feature selection method. The algorithms initialize the population in the dataset and perform selection, crossover, mutation and termination. The selection process is based on the concept of survival of the fittest. The formula to calculate the fitness function is given below

$$Fitness\ f(x) = \frac{fitness\ of\ an\ individual\ f(i)}{sum\ of\ fitness\ of\ all\ individuals\ f(I)}$$

The standard pseudocode for the Genetic Feature selection algorithm is as follows

**Algorithm 1.Genetic Feature Selection Algorithm**
*Initialize n = 0*
*Initialize the individuals in the population as P(n)*
*Evaluate fitness function for the individuals in the population P(n)*
*While {termination condition is not satisfied}*
*Do n = n+1 {iterate}*
*Selection () {pick individuals with better fitness}*
*Crossover () {combination of parent to form new individuals}*
*Mutation () {Making changes like bit flip}*
*End while*
*Return {fittest individuals in the population}*

## VI. Multi-Objective Evolutionary Fuzzy Classifier

The MOE Fuzzy is a fuzzy rule based classifier of three types: pre-niche, ENORA or NSGA II. The main objectives of this algorithm are to improve accuracy of classification, increase area under ROC and minimize RMSE to select best fittest non-dominant individuals from the population. This algorithm is implemented in [11] for survival prediction in an Intensive care unit. Table 1 provides information on the parameter configuration of the MOE Fuzzy classifier implemented in this paper.

Due to feature selection, the size of the dataset gets reduced and thus rules based classifiers may end up with the problem of model overfitting. Using MOE Fuzzy classifier, the model improves the classifier rate with minimum rules. The interpretability of the machine learning model on critical medical problems is improved. Hence, we have chosen MOE

Fuzzy classifier to analyses the performance of feature selection on Pima Indian Diabetes dataset.

MOE Fuzzy is an evolutionary computation technique that analogizes the nature of evolution. According to the theory of survival of fittest, a fitness function is formulated to evaluate the best suitable individuals from the population. The common characteristics of the MOE algorithms are Pittsburgh approach for variable length representation of the variables, repairing of constraints and adaptive variation.

In Niched preselection, the niched behavior of the evolution is induced. Among the search space of n niches, if the fitness of an offspring exceeds the fitness of the parent, the offspring replaces the parent and the evolution goes on. The niched-preselection algorithm should satisfy the following condition.

Where *Smin* is the minimum Niche selection rules and *Smax* is maximum niche selection rules*, C {P,i)* is the niche count that is the number individuals of the population *P*. The number niches n can be set to *Rmax –Rmin +1* for solving the MOE problem, where *Rmax* is Maximum Rule *s* and *Rmin* is Minimum rules. Thus the number of individuals will be within a range of *Smin* and *Smax*.

The ENORA algorithm applies a survival strategy optimization method to select the fittest population. The Population is initialized as P with i individuals. The individuals are evaluated with binary tournament algorithm. For binary tournament selection the population size must be equal to the number of children. The rank crowding algorithm is used for ranking the parents. Based on the ranks, the top two parents are selected for crossover. At the end of cross-over, two children are formed based on the variation algorithm.

$$Sample\ space\ of\ niches\ N(i) = \{1,2,3,\ldots.,n\} \quad (1)$$

$$S_{min} \leq C\{P,i\} \leq S_{max} \quad (2)$$

NSGA II is an improvement on the previous NSGA algorithm with an explicit diversity technique. It follows the same strategy as the ENORA, but ranking of population mechanism is quite different. For Ranking the NSGA II algorithm considers the entire population, whereas ENORA ranks a part of the entire population.

**Algorithm 2.Multi-Objective Evolutionary fuzzy algorithm**

*Require: T > 1 {No of iterations}*
*Require: I > 1 {No of individuals in population}*
*Require Amin, Rmax {Minimum accuracy of the classifier and maximum rules}*
*Initialization of Population P with i individuals*
*Evaluation of all individuals I in P*
*t =0 {initialize the iteration count as 0}*
*while t < T do*
*Q= Ø {initialize newly formed population}*
*i = 0*
*while i < I do*
*Select P1, P2← Binary tournament selection on P {Parent 1, Parent 2}*
*Select C1, C2← Variation (P1, P2) {Child1, Child 2 with variation algorithm)*
*Evaluate C1 and C2*
*Q ← Q Union {C1, C2}*
*i ← i + 2*
*end while*
*R = P Union Q*
*P= I { Rank-Crowding algorithm selection on R}*
*t = t + 1*
*end while*
*return S = si {Non-dominated individuals from P}*
*//Decision Making*
*Remove from S all solutions si such that A(si)<Amin*
*while S ≠ Ø do*
*Select si with max A (si) {most accurate solution}*
*If si is interpretable by Linguistic labelling algorithm*
*Return si*
*else*
*Remove si from S*
*end if*
*end while*
*return*

## VII. EXPERIMENTS

The experiment is made in Weka 3.8, a java package introduced by University of Waikato. The parameter configuration for both Feature selection and Classification is listed below

| **Parameters** | |
|---|---|
| ***a.*** | **Feature Selection: Genetic Search** |
| | Evaluator: CFS Subset Evaluation [12]<br>Crossing over Probability: 0.6<br>Maximum Generations: 20<br>Mutation Probability: 0.033<br>Probability Size: 20 |
| ***b.*** | **Classification Algorithm: Multi-objective Evolutionary fuzzy ENORA and NSGA II** |
| | Population size: 100<br>Evaluation metrics: Accuracy<br>No of Generations: 10<br>Max Rules: 1<br>Max Similarity: 0.1<br>MaxV: 2.0 MinV: 30.00<br>No of decimal places: 4<br>Population Size: 100<br>Report Frequency: 10<br>Seed: 1 |

## VIII. RESULTS AND EVALUATIONS

The results evaluated on the PID dataset with the supervised machine learning of 70% training set and 30% test set. The performance of the classifier is compared on original dataset and the feature selected dataset. Both ENORA and NSGAII fuzzy classifier are taken for the evaluation. The GA Feature selection is applied to the dataset and the classification procedure is repeated on the selected features and the accuracy of prediction is tabulated in Table 3. In addition to that the cost of the feature selection is calculated with the formula

$$\text{Cost of Feature Selection C (x)} = \frac{number\ of\ features\ selected\ from\ a\ dataset}{Total\ number\ of\ features\ in\ a\ dataset}$$

The Classification rate or accuracy of the machine learning algorithm is calculated with the following formula based on the True Positive (TP- Correctly classified as True), False Negative (FN - Wrongly classified as False), True Negative (TN- Wrongly classified as True) and False positive (FP- Correctly classified as False).

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

TABLE I. ACCURACY OF CLASSIFICATION ON THE SELECTED FEATURES FROM PID DATASET

| ***Algorithm*** | ***Accuracy before FS*** | ***Accuracy after GA FS*** |
|---|---|---|
| *Naïve Bayes* | 76.9565 % | 79.1304 % |
| *J48 graft* | 76.5217 % | 76.9565 % |
| *MLP* | 75.2174 % | 79.5652 % |

| | | |
|---|---|---|
| ***MOE ENORA*** | 80.4348 % | 81.3043 % |
| ***MOE NSGA II fuzzy (Proposed)*** | 78.2609 % | **83.0435 %** |

### *A. Evaluations:*

Multi-Objective evolutionary classifier is a good performer in terms of accuracy yet it is comparatively slower due to its complexity. The amount of features in a dataset highly influences the time taken for classification. Hence, in this paper we have chosen a genetic feature selection algorithm which has been proven as a better method on Pima Indians Diabetes dataset in the existing literature.

1. *Feature Selection*: The Genetic algorithm *{Algorithm 1}* based feature selection on Pima Indians diabetes dataset has resulted 4 features out of 8 existing features: Plasma glucose concentration, Body mass Index, Diabetes pedigree function and age in years. From the cost of feature selection formula C(x) is 0.5. The feature selection on Pima Indian Diabetes dataset reduces the dataset into half of its size.

2. *Classification Rate Analysis*: The effect of feature selection on Pima Indian diabetes dataset is evaluated with Naive Bayes, J48 Decision tree graft, Multi-Layer perceptron, Multi-Objective ENORA and Multi-Objective NSGA II. With Original set of features the MOE ENORA performs better than other classifiers with an accuracy of 80.4348 %. After feature selection the MOE NSGA II shows a drastic improvement to 83.0435 %. After feature selection with Genetic algorithm NSGA II shows an improvement of 5% approximately whereas ENORA improves its accuracy by 1%. The accuracy of the evolutionary classifiers before and after feature selection outperforms the classification rate exhibited by other state of art algorithms. In terms of classification rate the feature selection shows a positive impact on the Multi-Objective Evolutionary algorithms.
3. *Model Overfitting Problem*: The Rule based classifiers are prone to the problem of model overfitting. Hence performance of the Multi-Objective feature selection algorithms are testing with a test tests of range 10% to 90% and their classification rates are evaluated for consistency.

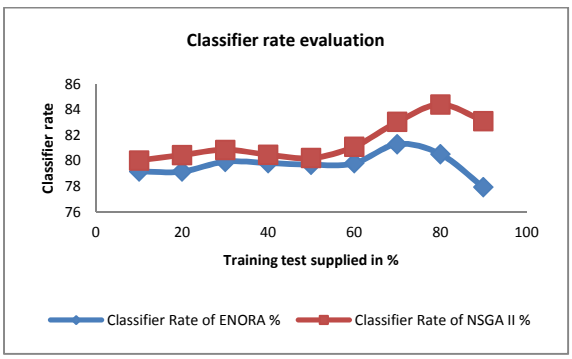


Fig.2. Evaluation of classifier rate of NSGA II and ENORA on different training rates to test model overfitting

According to Fig.2 the consistency of classifier rate is better in NSGA II than ENORA. As the number of training samples increase the performance of ENORA also improves whereas the NSGA II algorithm exhibits a consistency of above 80% accuracy. Suddenly both the algorithms drop a little at 90% training after a sustainable growth in other levels. From the fig.2 the fewer amounts of chances of model overfitting is observed in Multi-Objective evolutionary algorithm.

4. *Error rate evaluation:* The root mean square error observed at different levels of training datasets is represented in fig.3.

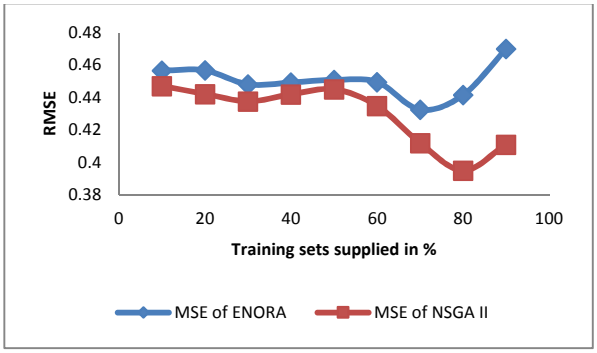


Fig.3. Evaluation of Root Mean square error rate of NSGA II and ENORA on different training rates

Any machine learning algorithm should possess reduced amounts of RMSE values to prove itself as a better performer. From the above analysis on RMSE it is clear that the NSGA II outperforms the ENORA at all stages of training with comparatively lesser RMSE rate. Hence, in terms of RMSE the NSGA II is a better performer than ENORA.

## IX. CONCLUSION

Medical data mining is a critical field where each and every decision matters. Rule based classifiers are not a very good choice in the case of critical decision support systems. Hence we have chosen a Multi-Objective Evolutionary fuzzy classifier that works on the principle of maximum classification rate and minimum rules. As the classifier is a slow performer, we have chosen the option of feature selection in the pre-processing stages that reduces redundancy and irrelevance among features, which positively induces the speed of performance and accuracy of classification. The Genetic feature selection algorithm adapted in many state of art literatures is implemented on Pima Indians diabetes dataset and the accuracy observed in the literature and the proposed work is compared. As a result the proposed method has shown a better performance than existing methods. Examining the features resulted by genetic algorithm, the feature reduction rate is 0.5, which is certainly a good performance but the important feature considered by medical experts for decision making on diabetes 'Plasma Glucose Concentration' is eliminated as an irrelevant feature. Hence, the reason behind this elimination has to be addressed and should overcome in future work. According to UCI repository after 2011, the

diabetes dataset is declared to have missing data in it. Hence the reason behind the error in feature selection may be due to the presence of outliers and missing values in the Dataset. In future work, the effect of outliers and missing data on feature selection should be analyzed and addressed.

## REFERENCES


[1] [1] K. G. M. M. Alberti and P. Z. Zimmet, "Definition , Diagnosis and Classification of Diabetes Mellitus and its Complications Part 1 : Diagnosis and Classification of Diabetes Mellitus Provisional Report of a WHO Consultation," pp. 539–553, 1998.

[2] [2] C. Weyer, C. Bogardus, D. M. Mott, and R. E. Pratley, "The natural history of insulin secretory dysfunction and insulin resistance in the pathogenesis of type 2 diabetes mellitus," vol. 104, no. 6, 1999.

[3] [3] J. P. Cunningham and Z. Ghahramani, "Linear Dimensionality Reduction: Survey, Insights, and Generalizations," vol. 16, pp. 2859–2900, 2014.

[4] [4] D. K. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection," in Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), 2017, pp. 451–455.

[5] [5] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995, pp. 338–345.

[6] [6] D. K. Choubey and S. Paul, "GA _ J48graft DT : A Hybrid Intelligent System for Diabetes Disease Diagnosis," vol. 7, no. 5, pp. 135–150, 2015.

[7] [7] D. K. Choubey and S. Paul, "GA _ MLP NN : A Hybrid Intelligent System for Diabetes Disease Diagnosis," no. January, pp. 49–59, 2016.

[8] [8] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco, and L. Miralles, "Multi-objective evolutionary feature selection for online sales forecasting," Neurocomputing, vol. 234, no. November 2016, pp. 75–92, 2017.

[9] [9] H. R. Kanan and K. Faez, "An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system," Appl. Math. Comput., vol. 205, no. 2, pp. 716–725, 2008.

[10] [10] D. Goldberg, "Genetic algorithms in search, optimization, and machine learning, 1989," Read. Addison-Wesley, 1989.

[11] [11] F. Jiménez, G. Sánchez, and J. M. Juárez, "Multi-objective evolutionary algorithms for fuzzy classification in survival prediction," Artif. Intell. Med., vol. 60, no. 3, pp. 197–219, 2014.

[12] [12] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.