International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Upgrading the Performance of KNN and Naïve Bayes in Diabetes Detection with Genetic Algorithm for Feature Selection

Ratna Nitin Patil*[1], Dr Sharvari Chandrashekhar Tamane[2]
*[1]Department of Computer Engineering, Research Scholar, Babasaheb Ambedkar Marathwada University, Aurangabad, India
[2]Department of Computer Engineering JNEC, Aurangabad, India.

ABSTRACT

Data mining is a science that is used for finding models and association rules in huge data where other statistical analysis cannot do that. The medical science needs data mining for analyzing massive data and producing predictive models. The purpose of this research is to present a framework for creating, evaluation and exploitation of data mining models. In this study we have used the combination of artificial intelligent technique such as feature selection with k Nearest Neighbor (kNN) and Naïve Bayes approach to develop a predictive model which classifies the patient as healthy and diabetic [1][2]. The main purpose of feature subset selection is to reduce the number of features used in classification while maintaining the acceptable classification accuracy. Our proposed Genetic algorithm (GA) for feature selection approach improves the classification accuracy and uses fewer input features [3]. Some attributes in the dataset may not be useful for diagnosis and thus can be eliminated before learning. The goal of feature selecting is to find a least set of attributes so that the resulting probability distribution of the data classes is close to the original distribution obtained by all attributes [4]. Irrelevant, redundant, or noisy data can be removed by using Genetic Algorithm which in turn improves the mining performance such as predictive accuracy and result comprehensibility

Keywords : Data Mining; Diabetes Mellitus; feature selection; Classification; Genetic Algorithm.

I. INTRODUCTION

According to the World Health Organization (WHO), diabetes is currently one of the biggest health concerns that the world is faced with. Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. A common effect of diabetes is Hyperglycemia or increased blood sugar. Un-monitored prevalence of diabetes also results in increased risk of vascular complications like Cardiovascular, renal, neural and visual disorders which are related to the duration of the disease. Diabetes causes some serious health issues including blindness, kidney failure, stroke and heart diseases.

Type 1 diabetes occurs when the body produces insufficient quantities of insulin. It is usually detected more in children. Type 2 diabetes occurs when the body does not effectively use the insulin produced. This is very frequently due to lack of physical activity, obesity, or incorrect dietary habits. Gestational diabetes occurs among pregnant women. In about 90 percent of cases, it is Type 2 diabetes that people are suffering from. The occurrence of Type 2 diabetes or Diabetes Mellitus may be prevented or delayed by adopting a healthy lifestyle.

Diabetes can be identified by symptoms – frequent urination, unusual thirst, excessive fatigue and

hunger, weight loss, and wounds that take long to heal. Type 2 diabetes, however, may remain unnoticed and patients may not display any signs for years. However, medical experts feel that timely detection and right management can go a long way in helping patients lead a normal life. Diabetes is a chronic medical condition, that is, it can be curbed at the initial level by introducing lifestyle changes and controlled after its incidence through medicines in early stages and administration of external insulin in advanced stages.

Numerous techniques have been developed for the diagnosis of diabetic mellitus. Most of the techniques used clustering and classification for the effective diagnosis of the diabetic mellitus disease. But, there is always a scope for improvement and still several techniques are being developed to overcome the limitations of the present techniques. This paper presents an analytical study on the existing techniques available for diabetes mellitus [5]. The characteristic features of the approaches are investigated to develop a better approach for the early and efficient diagnosis of the disease.

Genetic Algorithm (GA) is a computational concept of biological evolution that can be used to solve optimization problems [5]. The GA proposed by Holland, is a probabilistic optimal algorithm that is based on the evolutionary theories [6]. GA is based on a population of chromosomes. Successive populations of possible solutions are generated in a stochastic manner following laws similar to that of natural selection. The algorithm encodes a potential solution to a specific problem on a simple chromosome-like data structure and applies recombination operators to the structure so as to preserve significant information.

This paper is organized as follows: Section II briefs about the related work, Section III gives the methodology of the proposed system, Section IV gives the description of GA, Section V gives use of GA for Feature selection and optimization, Section VI describes classification algorithms, Section VII shows the experimental results and comparison and Section VIII concludes the paper.

II. RELATED WORK

Barakat et al. (2010) worked on the classification of diabetes disease using a machine learning approach such as SVM. A sequential covering approach for the generation of rules extraction is implemented using the concept of SVM which is an efficient supervised learning algorithm. The paper discusses eclectic rule extraction technique for the extraction of rules set attributes from the dataset such that the chosen attributes can be used for classification [7].

Aslam et al. (2013) implemented an expert system for the classification of diabetes data by Genetic Programming (GP). The technique implemented here consists of three stages: the first stage includes feature selection using t-test and Kolmogorov–Smirnov test and Kulback–Leibler divergence test, the next stage uses GP which is used for the non-linear combination of selected attributes from the first stage. At the final stage the generated features using GP is compared with K-Nearest Neighbor (kNN) and SVM. The selected features are then used for the classification with high accuracy.

Kala et al. (2011) proposed a new methodology for the diagnosis of breast cancer using the concept of neural networks. In this paper, a mixture of various expert models is congregated to solve various problems. The decision from each of the individual expert system is mixed to give a final output. The proposed architecture implemented here is used for the diagnosis of breast cancer by individually evolving neural network into Genetic Algorithm (GA). The experimental results performed on the methodology are highly scalable and provides efficient results on attributes and data items [8].

Parashar A. et al. (2014) have proposed Linear Discriminant Analysis and Support Vector Machine for the diagnosis of Pima Indians Diabetes dataset, where LDA reduces feature subsets and SVM is responsible to classify the data. They have also compared SVM with feed forward neural network (FFNN) but our proposed SVM+LDA gives better classification accuracy as 77.60% with 2 features [9].

Farahmandian M. et al. (2015) have applied diabetes

data set on various classification algorithms like SVM, KNN, Naïve bayes, ID3, CART and C5.0 to classify the diabetes data. They have compared the classification accuracy of these models. SVM gives best classification accuracy as 81.77% compare to others [10].

Jayalakshmi and Santhakumaran (2010) proposed a new and efficient technique for the classification of diagnosis of diabetes disease using Artificial Neural Network (ANN). The methodology implemented here is based on the concept of ANN which requires a complete set of data for the accurate classification of diabetes. The paper also implements an efficient technique for the improvement of classification accuracy of missing values in the dataset. It also provides a pre-processing stage during classification [11].

Patil et al. (2010) implements an association-rule-based technique for the classification of type-2 diabetic patients. The classification is implemented for Pima Indian Diabetes Dataset containing a number of attributes and classes. The methodology provides the generation of rules using apriori algorithm on the basis of some support and confidence. In the first stage, the numeric attributes are converted into categorical form which is based on the input parameters. Lastly generated the association rules which are useful to identify general associations in the data, to understand the relationship between the measured fields whether the patient goes on to develop diabetes or not. They have presented step-by-step approach to help the health doctors to explore their data and to understand the discovered rules better [12].

III. PROPOSED WORK

To establish a GA based feature selection, following are the main steps,

1. Data preprocessing: The main advantages of preprocessing is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Feature value scaling can help to increase accuracy as per our experiments.
2. Conversion of genotype to phenotype: This process converts each feature chromosome from its genotype into phenotype.
3. Feature subset: After applying the genetic operators and converting each feature subset chromosome from genotype into the phenotype a feature subset can be determined.
4. Fitness evaluation: Testing dataset is used to calculate the classification accuracy. After obtaining the classification accuracy each chromosome is evaluated by the fitness function. Fitness function is calculated by the equation (1).
5. Genetic operation: In this process system searches for better solutions by genetic operations selection, crossover, mutation and replacement.
6. Comparison of performance: In this process classification accuracy of unseen examples with the subset of selected feature is compared with the classification accuracy of unseen examples from validation set with all the features.

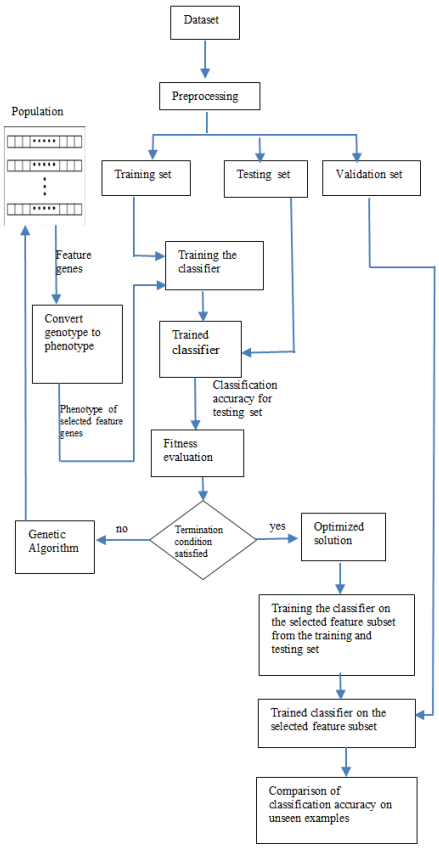


Figure 1. System Architecture of the proposed GA-based feature selection and parameter optimization

IV. GENETIC ALGORITHM

Genetic Algorithms are search and optimization techniques based on Darwin's Principle of Natural Selection. They can be used to solve Classification Problems. To use a genetic algorithm, you must represent a solution to your problem as a *genome* (or *chromosome*). The genetic algorithm then creates a population of solutions and applies genetic operators such as mutation and crossover to evolve the solutions in order to find the best one(s). Reproduction operators are applied in such a way that the chromosomes which provide a better solution to the objective problem are given more chances to reproduce themselves than those chromosomes which gives poorer solutions. The goodness of a solution is typically defined with respect to the current population.

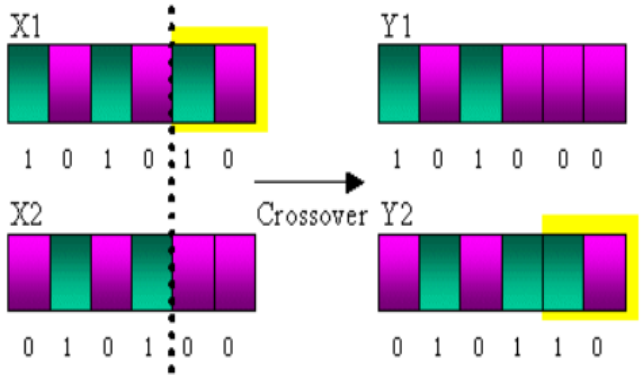


Figure 2. Genetic Crossover operator

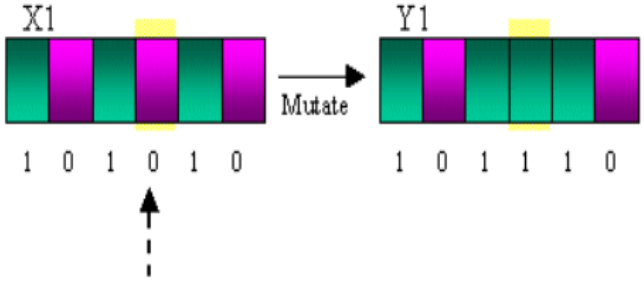


Figure 3. Genetic Mutation operator

GAs are population-based search techniques that maintain populations of potential solutions during searches. A string with a fixed bit-length usually represents a potential solution. In order to evaluate each potential solution, GA's need a payoff function that assigns scalar payoff to any particular solution. Once the representation scheme and evaluation function are determined, a GA can start searching. Initially, often at random, GAs create a certain number, called the population size, of strings to form the first generation. Next, the payoff function is used to evaluate each solution in this first generation. Better solutions obtain higher payoffs. Then, on the basis of these evaluations, some genetic operations are employed to generate the next generation. The procedures of evaluation and generation are iteratively performed until the optimal solution(s) is (are) found or the time allotted for computation ends [8]. Figure 4 represents the GA evolution flow.

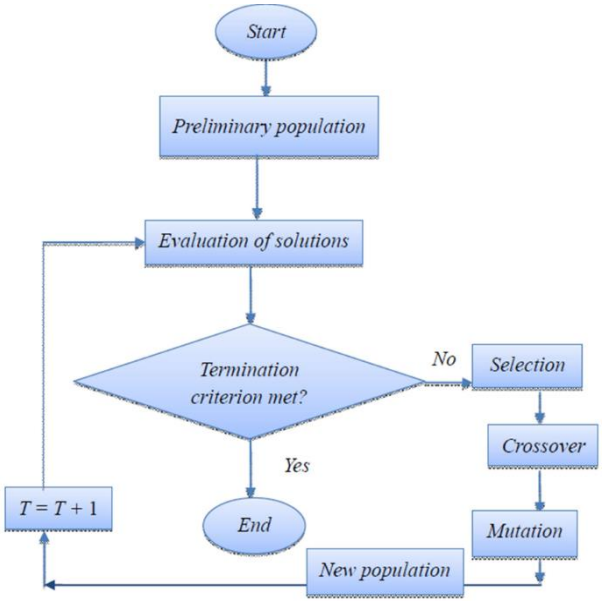


Figure 4. Schematic diagram of GA

V. GENETIC ALGORITHM BASED FEATURE SELECTION AND OPTIMIZATION

GAs search for the function optimum starting from a population of points of the function domain, not a single one. This characteristic suggests that GAs are global search methods. GAs use probabilistic transition rules during iterations, unlike the traditional methods that use fixed transition rules. This makes them more robust and applicable to a large range of problems [3].

Genetic Algorithm

1. Generate random population of N Chromosomes. These
   Chromosomes are potential solution for the given problem.
2. Evaluate the fitness function f(x) of each chromosome x in the random population.
3. Create a new population of Chromosomes by repeating the
   following steps until the new population is complete
   3.1 Select two parent chromosomes from the current population according to their fitness value (the better
   fitness, the higher chance to be selected)
   3.2 With a crossover probability the parents form a new offspring. If crossover operation was not done, offspring is an exact copy of parents.
   3.3 With a mutation probability alter new offspring in chromosome.
   3.4 Place new offspring in a new population.
4. Use newly generated population for the further generations.
5. If the termination condition is satisfied, stop the process and return the best solution in current population.
6. Go to step 2.

A. CHROMOSOME DESIGN

The chromosome is a representation of a solution candidate. In our GA-based feature subset selection, each individual is represented as a binary string encoding a feature subset. If the data consist of N features, an individual will be an N-bit binary string. For chromosomes representing the feature string, if a bit is having value '1' the feature is selected in the feature subset and the bit having value '0' represents the feature is not selected in the subset. Each individual in the population is thus a candidate feature subset [8].

B. FITNESS FUNCTION

The accuracy of kNN and Naive Bayes classifier is used as the fitness function for GA. The fitness function fitness(X) is defined as in equation (1).

$$\text{Accuracy}(X) = \text{fitness}(X) \quad (1)$$

Accuracy(X) is the test accuracy of testing data X by the kNN and Naïve Bayes classifier which is built with the feature subset selection of training data. The classification accuracy of is given by equation (2).

$$\text{Accuracy}(X) = (c/t) * 100 \quad (2)$$

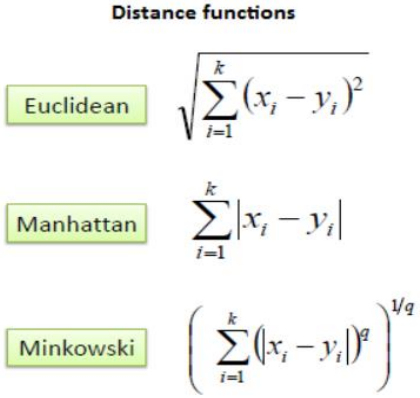
Where,
c - Samples that are classified correctly in test data
t - Total number of Samples in test data

C. CLASSIFICATION ALGORITHMS

A. k-Nearest Neighbor Classifier

In pattern recognition, the kNN algorithm is a method for classifying objects based on closest training examples in the feature space. kNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification [1].
In kNN, the training samples are mainly described by n-dimensional numeric attributes. The training samples are stored in an n dimensional space. When a test sample (unknown class label) is given, k-nearest neighbor classifier starts searching the 'k' training samples which are closest to the unknown sample or test sample. Closeness is mainly defined in terms of Euclidean distance. To find the closeness normally some distance measures are used. Sometimes one minus correlation value is also taken as a distance metric. For continuous variables the following three distance measures are used. They are Euclidean distance, Manhattan distance and Minkowski distance [13]. In the instance of categorical variables, the Hamming distance must be used. The distance measures between two points x (x_1,

$x_2, \ldots x_n$) and y ($y_1, y_2, \ldots y_n$) are calculated as given below,



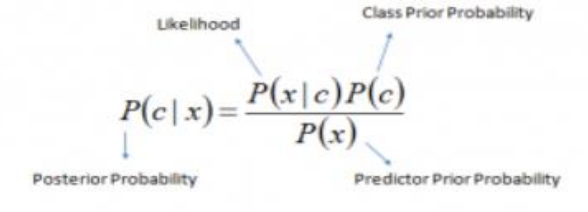
Pseudocode of k-Nearest Neighbor

Classify (X, Y, x) // X: training data, Y:class labels
// x: unknown sample
for i = 1 to m do
Calculate the distance $d(X_i, x)$
end for
Compute set I containing indices for the k smallest distances $d(X_i, x)$.
return majority label for $\{Y_i$ where $i \in I\}$

kNN classifier is one of the most popular neighbourhood classifier in pattern recognition. However, it has limitations such as: great calculation complexity, fully dependent on training set and no weight difference between each class. To combat this, a novel method to improve the classification performance of kNN using Genetic Algorithm is proposed in this paper.

B. Naïve Bayes

This classification technique is based on Bayes theorem with an assumption of independence among attributes. This classifier assumes that the presence of a particular feature is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods [14]. Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c) using the equation below:



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

where,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

VI. EXPERIMENTAL RESULTS AND COMPARISON

In order to assess the performance of the proposed method, PIMA dataset was analyzed [15]. PIMA Indian Diabetes Dataset from UCI repository contains 768 instances and 9 attributes. The PIMA diabetes dataset only represents the Indian national females who are at least 21 years old.

In this experiment, scaling is done on the PIMA Dataset and then the dataset is divided into training data, test data and validation data. Fitness function of the genetic algorithm is implemented on the training and test data on the classification accuracy. Validation data is left out to see if the classifier generalizes well by trying against the foundation data. The GA is configured to have population size as 100 and was run for 10 - 200 generations. In a conventional GA method, Crossover probability and Mutation probability is taken as 0.6 and 0.1.

TABLE I. GA PARAMETERS AND THEIR VALUES

Parameter	Value
Population size	100
Maximum no. of Generations	50
Selection Method	selTournament
Crossover Type	Single point
Crossover Probability	0.5
Mutation Type	Flip bit
Mutation Probability	0.2

Figures (5) - (7) show the results obtained from GA with kNN method. We have got the optimal feature subset of 5 features. The input dataset contains 8 features and we are able to drop 3 features. The validation accuracy (on unseen examples) has been improved from 75% to 83% using k Nearest Neighbor algorithm. The output after each generation of kNN algorithm is mentioned below.

```
Test accuracy (with all features(KNN):  0.707317073171
Validation accuracy for unseen examples with all features(KNN):0.74675324675
```

gen	nevals	avg	std	min	max
0	100	0.66878	0.0561814	0.528455	0.796748
1	65	0.706341	0.0470461	0.585366	0.772358
2	55	0.732846	0.0350447	0.634146	0.772358
3	70	0.737154	0.0362324	0.601626	0.772358
4	65	0.754228	0.0262403	0.650407	0.788618
5	64	0.753659	0.0332141	0.609756	0.788618
6	56	0.759268	0.0279253	0.609756	0.788618
7	58	0.763415	0.0200963	0.691057	0.788618
8	77	0.75935	0.0277139	0.609756	0.788618
9	55	0.760163	0.0252818	0.691057	0.788618
10	58	0.768537	0.021738	0.715447	0.796748

```
---Optimal Feature Subset(s)---

Percentile:                     0.5751633986928104
Validation Accuracy:            0.831168831169
Individual:     [1, 0, 0, 1, 1, 1, 0, 1]
Number Features In Subset:      5
Feature Subset: ['preg', 'skin', 'test', 'mass', 'age']
```

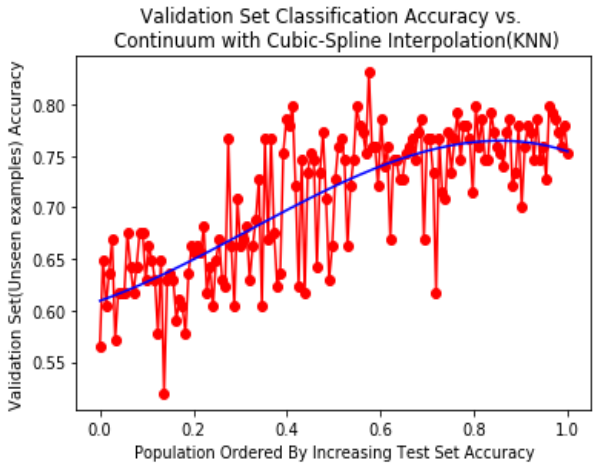


Figure 5. Validation set accuracy vs. Population ordered by Test set accuracy with Cubic-Spline Interpolation (kNN)

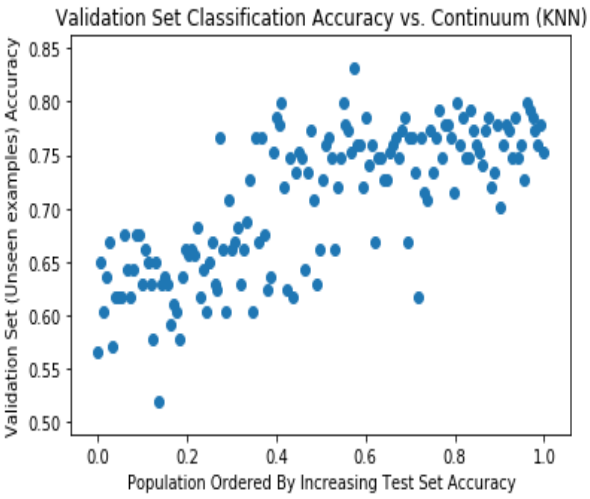


Figure 6. Validation set accuracy vs. Population ordered by Test set accuracy (kNN)

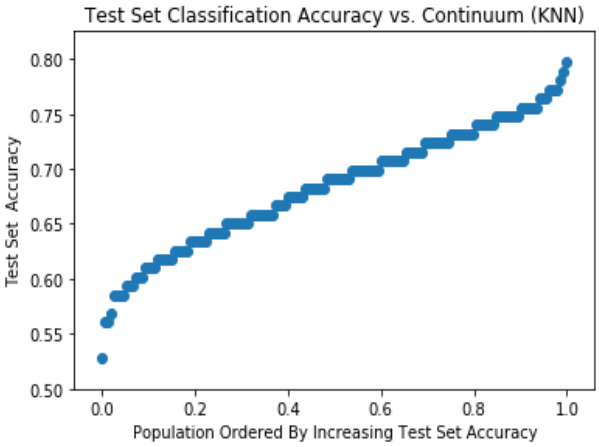


Figure 7. Test Set accuracy vs. Population ordered by Test set accuracy with Cubic-Spline Interpolation (kNN)

Figures (8) -(10) show the results obtained from GA with Naïve Bayes method. We have got the optimal feature subset of 4 features. The input dataset contains 8 features and we are able to drop 4 features. The validation accuracy (on unseen examples) has been improved from 75% to 81% using Naïve Bayes classification algorithm. The output after each generation of Gaussian Naïve Bayes algorithm is mentioned below.

```
Test accuracy (with all features(Naive Bayes):  0.756097560976
Validation accuracy for unseen examples with all features(Naive Bayes):0.7467532467

gen     nevals  avg             std             min             max
0       100     0.687236        0.0522158       0.569106        0.780488
1       71      0.726016        0.042206        0.569106        0.780488
2       69      0.75187         0.0216779       0.626016        0.780488
3       52      0.764797        0.0109231       0.723577        0.780488
4       53      0.769268        0.00953545      0.739837        0.780488
5       53      0.771057        0.0191953       0.650407        0.780488
6       62      0.774959        0.0144925       0.658537        0.780488
7       58      0.775041        0.0153422       0.650407        0.780488
8       51      0.774309        0.0188968       0.658537        0.780488
9       57      0.777886        0.00707644      0.747967        0.780488
10      50      0.777724        0.00691965      0.747967        0.780488

---Optimal Feature Subset(s)---

Percentile:                     0.8246753246753247
Validation Accuracy:            0.818181818182
Individual:     [0, 0, 0, 1, 0, 1, 1, 1]
Number Features In Subset:      4
Feature Subset: ['skin', 'mass', 'pedi', 'age']
```

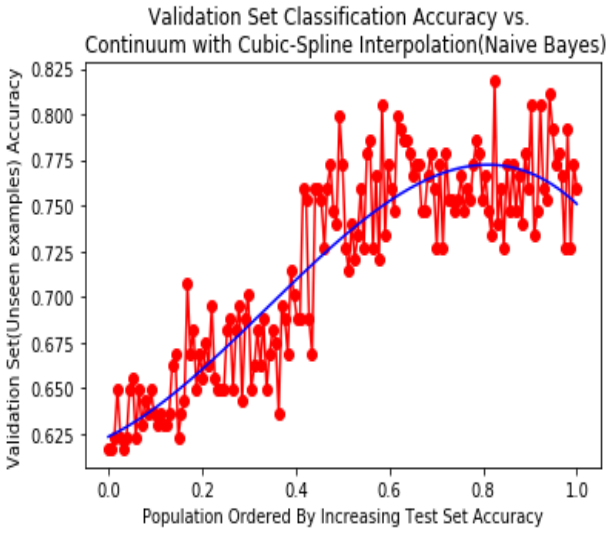


Figure 8. Validation set accuracy vs. Population ordered by Test set accuracy with Cubic-Spline Interpolation (Naive Bayes)

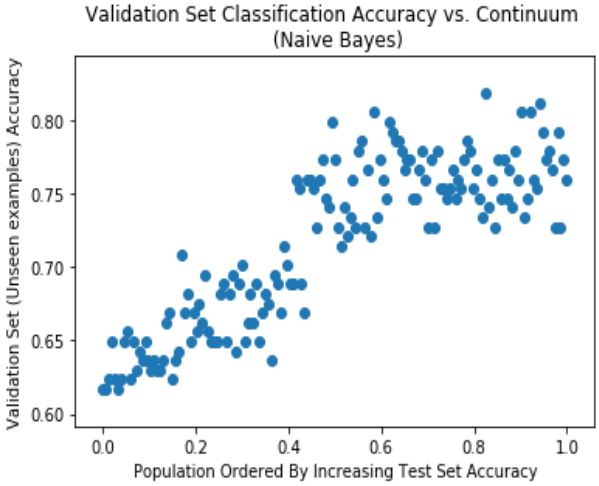


Figure 9. Validation set accuracy vs. Population ordered by Test set accuracy (Naïve Bayes)

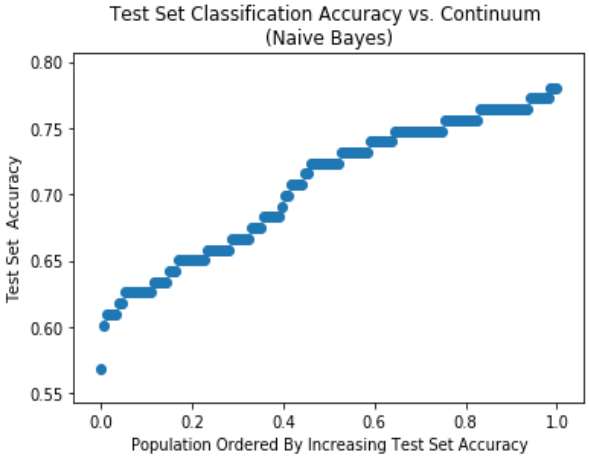


Figure 10. Test Set accuracy vs. Population ordered by Test set accuracy with Cubic-Spline Interpolation (Naive Bayes)

TABLE II. COMPARISON OF CLASSIFICATION ACCURACY WITH OTHER METHODS

Reference	**Methodology**	**Classification Accuracy %**
This Work	GA+kNN	83.12%
This Work	GA+ Naïve Bayes	81.82%
This Work	kNN	74.68%
This Work	Naïve Bayes	74.68%
Recognition of diabetes disease using a new hybrid	PSO	82.32%

learning algorithm for Nefclass (Daho et al., 2013) [16]		
A hybrid intelligent system for medical data classification (Seera and Lim, 2014) [17]	Fuzzy min–max neural network, CART	78.39%
Feature selection using fuzzy entropy measures with similarity classifier (Luukka, 2011) [18]	Fuzzy entropy measure	75.97%
Region-based support vector machine algorithm for medical diagnosis on Pima Indian diabetes dataset (Karatsiolis and Schizas, 2012) [19]	Modified support vector machine	82.2% (Not implemented for the real application framework)
A comparative study on diabetes disease diagnosis using neural networks (Temurtas et al., 2009) [20]	Probabilistic Neural Network (PNN)	79.62%
Design of a diabetic diagnosis system using rough sets (Margret et al., 2013) [21]	Rough set	76%
An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA– ANFIS (Dogantekin et al., 2010) [22]	LDA–ANFIS	84.61%
Feature generation using genetic programming with comparative partner selection for diabetes classification (Aslam et al., 2013) [23]	GPKNN, GPSVM	80.5%(GPKNN) 87.0% GPSVM) (Issues-Missing values are ignored)
Using fuzzy Ant Colony Optimization for diagnosis of diabetes disease (Ganji and Abadeh, 2010) [24]	ACO	79.48 (strong rules of have been diluted to improve the cooperation between rules)

VII. CONCLUSION

In this paper the model was developed and its accuracy was compared with the earlier models. The accuracy of the model outperformed with the earlier models. The accuracy can be further improved by using better fuzzy membership functions. In this paper, several existing techniques for the classification of medical diagnosis of diabetes patients have been discussed on the basis of advantages, issues and accuracy. More efficient techniques for the classification of diabetes patients can be implemented based on the discussion.

VIII. ACKNOWLEDGEMENT

We are thankful to the Department of Computer Science and Engineering, Babasaheb Ambedkar Marathwada University, Aurangabad, India for providing the research facilities. We are also thankful to all the members of National Institute of Electronics and Information Technology, Government of India, Aurangabad for their support.

IX. REFERENCES

[1] Nirmala Devi M.,Appavu S. , Swathi U.V., —An amalgam KNN to predict diabetes mellitus:, IEEE, 2013.

[2] Rashedur M. Rahman, Farhana Afroz " Comparison of Various Classification Techniques Using Different Data Mining Tools

for Diabetes Diagnosis", Journal of Software Engineering and Applications, 2013, 6, 85-97.

[3] C. Gunavathi, K. Premalatha "Performance Analysis of Genetic Algorithm with kNN and SVM for Feature Selection in Tumor Classification", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:8, No:8, 2014.

[4] Devi kalyan Karumanchi, James Dillon and Elizabeth Gaillard " Early diagnosis of Diabetes mellitus through the eye "2nd International Conference on Endocrinology October 20-22, 2014.

[5] Selvakuberan, K., Kayathiri, D., Harini, B. and Devi, M.I "An Efficient Feature Selection Method for Classification in Health Care Systems Using Machine Learning Techniques", IEEE. pp.8610–8615, 2011.

[6] Amir Amiri and Vahid Rafe "Hybrid Algorithm for Detecting Diabetes ", International Research Journal of Applied and Basic Sciences © 2014 Available online at www.irjabs.com ISSN 2251-838X / Vol, 8 (12): 2347-2353 Science Explorer Publications.

[7] M Nahla H. Barakat, Andrew P. Bradley, Senior Member, IEEE, and Mohamed Nabil H.Barakat "Intelligible Support Vector Machines For Diagnosis Of Diabetes Mellitus" IEEE Transactions On Information Technology In Biomedicine, Vol. 14, No. 4, July 2010 Digital Object Identifier10.1109/TITB.2009.2039485 July 2010.

[8] R.R.Janghel, Anupam Shukla, Ritu Tiwari, Rahul Kala "Breast Cancer Diagnostic System using Symbiotic Adaptive Neuro-evolution (SANE)" Proceedings of the 2010 IEEE International Conference of Soft Computing and Pattern Recognition, Cercy Pontoise/Paris, France, pp 326-329.

[9] Parashar, Rawat "Diagnosis of PIMA Indian Diabetes by LDA-SVM Approach", International Journal of Engineering Research & Technology, Vol 3, issue10, Oct2014, ISSN2278-0181.

[10] Norul Hidayah Ibrahim et.al " A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients" nternational Journal of Engineering and Technology, Vol 5 No 5 Oct-Nov 2013.

[11] Jayalakshmi, T. and Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks", International Conference on Data Storage and Data Engineering, DSDE 2010, Bangalore, India, pp.159–163.

[12] Patil, B.M., Joshi, R.C. and Toshniwal "Association rule for classification of type-2 diabetic patients", Second International Conference on Machine Learning and Computing 2010, IEEE.

[13] M. Kothainayaki, P. Thangaraj "Clustering and Classifying Diabetic Data Sets Using K-Means Algorithm" Article can be accessed online at http://www.publishingindia.

[14] Md. Mozaharul Mottalib, Md. Mokhlesur Rahman, Md. Tarek Habib and Farruk Ahmed "Detection of the Onset of Diabetes Mellitus by Bayesian Classifier Based Medical Expert System" Transaction on Machine Learning and Artificial Intelligence DOI: 10.14738/tmlai.44.1962 Publication Date: 19th July, 2016.

[15] K. Saravananathan and T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining," In proceeding of Indian Journal of Science and Technology, vol 9, no. 43, Nov 2016.

[16] Mostafa EL HABIB DAHO, Nesma SETTOUTI, Mohammed El Amine LAZOUNI, M. Amine CHIKH "Recognition of diabetes disease using a new hybrid learning algorithm for NEFCLASS" Conference: Conference: IEEE Proceeding of the 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), 2013.

[17] Seera, Lim "A hybrid intelligent system for medical data classification" Volume 41, Issue 5, April 2014, Pages 2239-2249.

[18] Pasi Luukka "Feature selection using fuzzy entropy measures with similarity classifier" Expert Systems with Applications Elsevier Volume 38, Issue 4, April 2011, Pages 4600-4607.

[19] Savvas Karatsiolis, Christos N. Schizas "Region-based support vector machine algorithm for medical diagnosis on Pima Indian diabetes dataset" pp:139-144DOIBookmark:.http://doi.ieeecomputersociety.org/10.1109/ BIBE.2012.6399663

[20] HasanTemurtas, NejatYumusak,FeyzullahTemurtas "A comparative study on diabetes disease diagnosis using neural networks" Expert Systems with Applications Elsevier Volume 36, Issue 4, May 2009, Pages 8610-8615.

[21] Margret Anouncia S., Clara Madonna L. J., Jeevitha P., Nandhini R. T. "Design of a Diabetic Diagnosis System Using Rough Sets" CYBERNETICS AND INFORMATION TECHNOLOGIES Volume 13, No 3 Print ISSN: 1311-9702; Online ISSN: 1314-4081 DOI: 10.2478/cait-2013-0030.

[22] EsinDogantekin, AkifDogantekin, DeryaAvci, LeventAvcid "An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS" Digital Signal Processing Elsevier Volume 20, Issue 4, July 2010, Pages 1248-1255.

[23] Muhammad Waqar Aslam et al. "Feature generation using genetic programming with comparative partner selection for diabetes classification" Expert Systems with Applications 40(13):5402-5412 · October 2013.

[24] Mostafa Fathi, Mohammad Saniee Abadeh "Using fuzzy ant colony optimization for diagnosis of diabetes disease" DOI: 10.1109/IRANIANCEE.2010.5507019 · Source: IEEE Xplore Conference: Conference: Electrical Engineering (ICEE), 2010 18th Iranian Conference.