

PAPER • OPEN ACCESS

## An efficient method for high performance prediction mechanism for diabetes using enhanced Firefly algorithm and Map-Reduce

To cite this article: R Sujatha *et al* 2021 *J. Phys.: Conf. Ser.* **1964** 042056

View the [article online](#) for updates and enhancements.



**ECS** **240th ECS Meeting**  
Digital Meeting, Oct 10-14, 2021

**Register early and save  
up to 20% on registration costs**

Early registration deadline Sep 13

**REGISTER NOW**

# An efficient method for high performance prediction mechanism for diabetes using enhanced Firefly algorithm and Map-Reduce

R Sujatha<sup>1\*</sup>, B Gomathi<sup>2</sup> and T Padma<sup>3</sup>

<sup>1</sup>Department of Computer Science, PSG College of Arts & Science, Coimbatore-641 014, Tamil Nadu, India.

<sup>2</sup> Department of Computer Science, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai-600 017, Tamil Nadu, India.

<sup>3</sup> Department of Computer Applications, Sona College of Technology, Salem-636 005, Tamil Nadu, India.

E-mail: sujamca2k5@gmail.com; gomathiphd2k15@gmail.com; padmatheagarajan@gmail.com

**Abstract.** Algorithms in Data mining are utilized to predict unrefined data into useful conventional information. This conservative information plays a vital role in the Health care industry. In this study we focus on the functionalities of diabetes prediction. In diabetes data we have the problem of data imbalance in predicting the accuracy. The Proposed tailored Firefly Algorithm along with Map reduce is used to augment the efficacy and precision of prediction. Comparison of Different bench mark algorithms with our new Extended Fire Fly is done and variety of classification methods are used with moto to increase the effectiveness. The new method helps to maximize the prediction of accuracy and reduces the time. The PIMA Indian Diabetic Dataset from UCI machine learning repository is utilized for our experiment results. Different metrics are used in order to prove the effectiveness.

## 1. Introduction

Data mining is surfing large data and finding out hidden patterns. Data mining is confluence of various techniques and is widely used in various applications. Health care industry is predominant field and data mining plays a vital role for systematic use of data and analysis. The application of DM in health care helped in policy decision and saved hospitals from error. DM techniques helped to prevent disease as the helped easy prediction. The applications of Data mining to health care data leads to best practices to improve care at reduced cost. Map Reduce framework integrated with Data mining techniques provides efficient method to handle data with very high speed considering security and also accuracy.

MapReduce principle splits the process into map phase and reduce phase. Key-value-pair is used as input and output by the phases. Enormous health care data is available and The phenomenon used in Map Reduce is storing and processing data. The key concept is to find out hidden knowledge in medical area using various mining techniques. SVM is used to classify the medical data set as they produce better performance in classification and SVM provide precise and accurate classification for health care data[1].



FireFly is also nature inspired algorithm and this is used in this paper because of its capability of automatic subdivision and FA has the capacity to deal with multimodal function.[2]. The vast availability of medical data leads to improvement in data mining tools to ensure health care professionals to diagnose diabetes disease. Data mining tools and techniques are used to diagnose variety of disease like stroke, cancer, cardiac disease. Research confirms that diabetes is leading cause for death. Hybridizing data mining with other techniques like Map reduce and optimization results in performance enhancements. Gap analysis in prediction of diabetes disease is analyzed in order to give treatment in proactive manner. The focus of this paper is to fuse the advantages of Mapreduce, SVM with enhanced Firefly algorithm to increase the efficiency to predict diabetes.

## 2. Literature Review

Existing technologies does not support unstructured healthcare data and in the paper “Health-care Big Data Analysis using Hadoop MapReduce” the author suggested for storing an processing the usage of MapReduce.[3]. Utilization of various models with respect to medical data was studied by Neesha Et al. [4].

In the research work [5] “A MapReduce based parallel SVM for large scale spam filtering”, the author represents a parallel algorithm for filtering. Ontology also employed to maximize accuracy. Firefly is used all in areas of optimization and also in engineering practice. This algorithm produces better results and the author proposed a comprehensive review of the algorithm[6].

A. Iyer et al [7] attempted to discover solutions to analyze the diabetes ailment with the support of studying the styles discovered by implementing classifier Naïve Bayes.

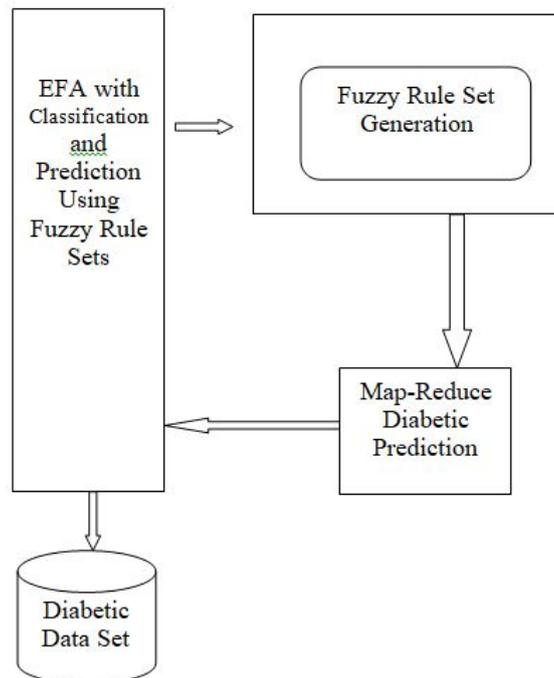
The important to do step is preprocessing which in turn contributes to Feature Selection [9]. It permit us to decrease variety of attribute wanted for closing the end outcome, eliminate unneeded information, inappropriate information and missing information. This pace up the prediction rules and allow in generating better accurate consequences which is primarily based at the preprocessed records set [8]. The author iterates that performance and exactness of classifier is stepped forward while it is fixed with any feature choice [9]-[10].

Various strategies and methods are available for Healthcare system to diagnose diabetes in minimum possible time with maximum accuracy. A mechanism for identifying category 2 diabetes mellitus was illustrated by N. Wang & G. Kang [11]. They used hybrid mechanism and arrived at better solution.

Based on the prediction rules the author suggested mobile application in order to pro-actively review the stages of diabetes[12]. Android mobile application by using predication rule was developed by Sowjanya [13] for monitoring diabetes. Shi et al. [14] work includes assessment of the various risk involved due o diabetes using gadgets. Researchers had proved hybridization yields better results than used individually[15].

## 3. Proposed System

The proposed model is controlled to upgrade the accuracy to anticipate the diabetic patient. during this model, we've differentiated different classifiers and ACO, PSO count. the object is to grow the precision by using Extended Firefly Algorithm (EFA) close by Fuzzy rule set and Map lessen technique. PIMA Indian Diabetes Dataset from UCI AI vault is utilized for our execution work.



**Figure 1.** The proposed model using Classification for Diabetes Datasets.

3.1. Fuzzy rule set generation

The rules set are created based on the data models. In this step, the technique reads the data. Using this, the lists of features are formed. For every feature values, the model calculates the minimum maximum values. Based on the result, the values are and the fuzzy rule has been created. This is applied to the dataset to frame the rule set.

**Algorithm**

**Input:**Feed the Cluster set Fcs

**Output:**Fuzzy Rule Set Rs is generated

Start

Read Fcs.

Identify number of classes  $NEs1 = \sum \text{Distinct (Ensembles Class)} \in Fc$

For each class c1

Read all data arguments Cds1.

$Cds1 = \sum Di1 \in c1$

Find list of features  $Fs1 = \sum \text{Features} \in Di1$

For each feature  $Fi1$

Calculate Feature minimum  $Fmin1 = \int_{i=1}^{size(Cds)} \min(Cds1(1(Fi1)))$

Calculate Feature maximum  $Fmax2 = \int_{i=1}^{size(Cds)} \max(Cds1(1(Fi1)))$

End

Create Fuzzy Rule  $RS = \int_{i=1}^{size(Fs1)} \sum Fmin1, Fmax1$

Insert to rule set Rs.

End  
Stop.

The above algorithm calculates the minimum and maximum values for each and every feature in the dataset. The range values are measured based on the feature and finally the fuzzy rule is generated.

### 3.2. Firefly algorithm (FA)

Firefly algorithm (FA) is an effective optimization technique. Like PSO, Firefly algorithm (FA) is a population based search algorithm. Each member represents a solution in the search space. Based on this an Extended Firefly algorithm (EFA) is proposed to find out the accuracy of the diabetic data [25]. The below is the Pseudo Code for code of Extended Firefly algorithm [24].

#### Pseudo Code for code of Extended Firefly algorithm

Begin

Let fitness function be  $f(ax)$  where  $x1 = (x_1, x_2, \dots, x_D)$

Generate an initial population of fireflies  $xi1 (i = 1, 2, \dots, n)$

Light intensity  $Ii1$  at  $xi1$  is determined by  $f(xi1)$

Define the light absorption coefficient  $\gamma1$

**While** ( $t1 < MaxGeneration$ )

**for**  $i1 = 1$  to  $n$  (all  $n$  fireflies)

**for**  $j1 = 1 : n$  ( $n$  fireflies)

**if** ( $Ij1 > Ii1$ )

                move fireflies  $i1$  towards  $j1$  in  $d$ -dimension

**endif**

                Attractiveness varies with distance  $r$  via  $\beta_0 e^{-\gamma r^2}$

                Evaluate new solutions and update light intensity;

**end for**  $j1$

**end for**  $i1$

Rank fireflies and find the current best;

**end while**

Post-processing the results and visualization;

**end**

The limit of the classifier is to predict diabetes and normal individual from the dataset stacked. one among the subset is planning set and another is test set. The past holds set of segments named with class marks and thusly the test set contains lines with dark names.

The above is the procedure for Extended Firefly algorithm [23] used to predict and reduce the features of diabetic data.

The manners of bird are used in particle swarm optimization (PSO) (Kennedy and Eberhart, 1995) [22], and ant actions in ant colony optimization (ACO) (Dorigo et al., 2006) [20], artificial

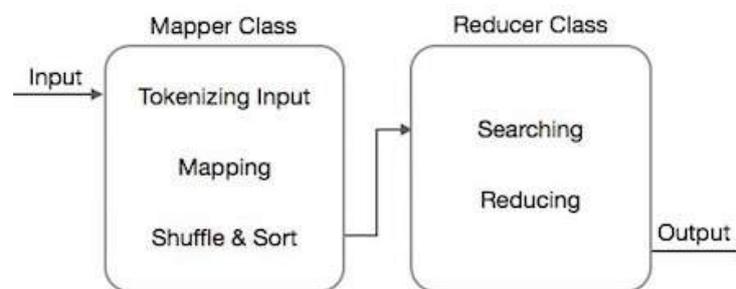
bee colony (ABC) algorithm motivated by the behavior of bee colonies (Karaboga and Basturk, 2007) [21].

### 3.3. MapReduce

The MapReduce algorithm has two main tasks, Map and Reduce.

- (i) The mapping job is executed by Mapper Class
- (ii) The reduce task is executed by Reducer Class.

Identification of input and sorting is done Map class the output is utilized as an input to the Reducer class, and this is used to search similarity leading to reduction.



MapReduce generate several mathematical procedures to crack a task into small portions and allocate them to various system. MapReduce is used in our diabetic data after the process of fuzzy rule generation along with the Extended Firefly Algorithm (EFA) algorithm to predict the disease and to improve its efficiency and accuracy.

## 4. Results and Discussion

The proposed EFA algorithm fused with fuzzy rule for prediction proves to be effective when applied to varied data. This integration proved good for predicting diabetes data.. The classification accuracy is also improved to obtain optimized results. The performance is compared with few existing techniques such as particle swarm optimization (PSO) and ant colony optimization (ACO).

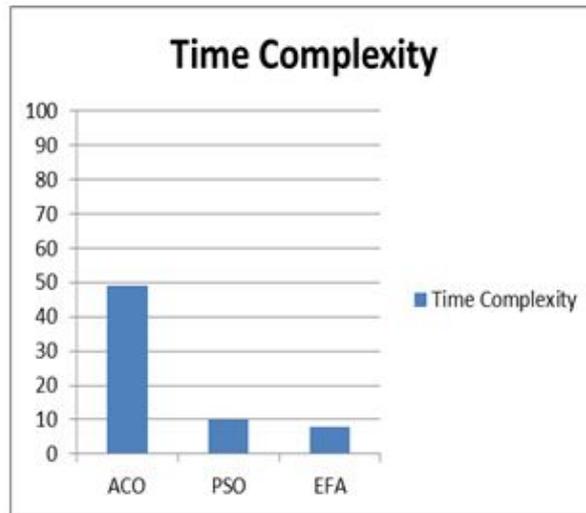
### *Time Complexity*

The time complexity of collection and gauge has been assessed for the proposed estimation EFA differentiating and hence the current computation particle swarm improvement (PSO) and bug territory smoothing out (ACO). It's been gone after for different number of trials.

**Table 1.** Comparison on classification time complexity.

Method Name	Time Complexity
ACO	49
PSO	10
EFA	8

Table 1 shows the close to result on time unpredictability introduced by different procedures and in this way the proposed estimation has conveyed less time multifaceted nature than various systems.



**Figure 2.** Comparison of time complexity.

Figure 2 shows the near outcome on time intricacy passed on by different strategies and subsequently the outcome shows that the proposed technique has made less time than different systems.

*False Classification Ratio*

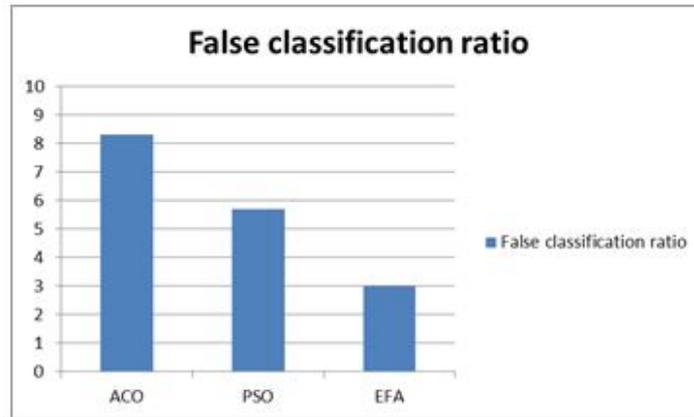
The False Classification Ratio has been measured for the proposed algorithm Extended Firefly algorithm (EFA) comparing with the existing algorithm particle swarm optimization (PSO) and ant colony optimization (ACO). It has been tested for different number of test cases.

The procedure used to lessen the higher number of highlights by the utilizing the bogus depiction degree. By including the comfortable principles, the bogus solicitation degree is decreased and helps the expansion of figure precision.

**Table 2.** Comparison on false classification ratio.

Method Name	False classification ratio
ACO	8.3
PSO	5.7
EFA	3

Table 2, shows the overall result on counterfeit solicitation degree for various techniques and the proposed calculation EFA has made inconsequential inspiration for bogus degree than the other existing strategies.



**Figure 3.** Comparison on false classification ratio.

Figure 3 shows the near outcome on bogus arrangement degree made by different methodology and its shows evidently that the proposed procedure has passed on less fake depiction degree than different techniques.

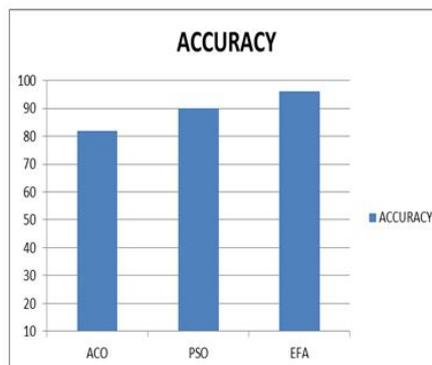
*Prediction accuracy*

The presumption accuracy has been broadened. The proposed assessment join the highlights so the accuracy will be broadened

**Table 3.** Comparison on prediction accuracy.

Method Name	Accuracy
ACO	82
PSO	90
EFA	96

Table 3 shows the comparative result on prediction accuracy introduced by different methods and the proposed algorithm has produced higher prediction accuracy than other methods.



**Figure 4.** Comparison on disease prediction accuracy.

Figure 4 shows the association on infection gauge of various techniques. The proposed system produces higher illness suspicion accuracy than different procedures.

#### *Comparison on prediction method with Presion, Recall and Accuracy*

From the above results, it is clearly shown that the proposed system provides the best result when predicting the information tic data. To evaluate the efficiency, we utilized evaluation metrics: precision, recall, and F-measure which is discussed below.

**Table 4.** Comparison on prediction accuracy.

Method Name	Presion	Recall	Accuracy
ACO	75	80	82
PSO	83	86	90
EFA	94.5	96	96

## 5. Conclusion

The core objective of this research is to enhance the accuracy of predictive model for diabetic data. The proposed EFA algorithm along with fuzzy set and map reduce is used to improve the accuracy for diabetic data set. The model is also utilized for prediction has been applied and evaluated to improve the efficiency with different test cases. The proposed method has produced effective results in predicting diabetic data. The classification accuracy is also improved to obtain the best results. The performance of the proposed system is compared with few existing techniques such as particle swarm optimization (PSO) and ant colony optimization (ACO). The PIMA diabetic dataset is used. The dataset used for the used for the evaluation of the EFA along with map reduce proposed method contains the parameter data set name, no of dimensions, no of data points and no of classes. For best results Precision and Recall is also calculated for the model.

## References

- [1] Padmavathi Janardhanan, Heena L and Fathima Sabika 2015 Effectiveness of support vector machines in medical data mining *Journal of Communications Software and Systems* **11(1)**.
- [2] Firefly Algorithm: Recent Advances and Applications <https://www.researchgate.net/.../255971821>, AUG 2013
- [3] Thanuja Nishadi A S 2019 Healthcare big data analysis using Hadoop MapReduce *International Journal of Scientific and Research Publications* **9(3)**.
- [4] Neesha Jothia , Nur'Aini Abdul Rashidb and Wahidah Husainc 2015 Data mining in healthcare –A review. <https://www.sciencedirect.com/science/article/pii/S1877050915036066>.
- [5] Godwin Caruana, Maozhen Li and Man Qi 2011 A MapReduce based parallel SVM for large scale spam filtering *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*.
- [6] Iztok Fister 2013 A comprehensive review of firefly algorithms *Swarm and Evolutionary Computation* **13**, 34-46.
- [7] Aiswarya Iyer, Jeyalatha S and Ronak Sumbaly 2015 Diagnosis of diabetes using classification mining techniques *International Journal of Data Mining & Knowledge Management Process (IJDMP)* **5(1)**.
- [8] Jain A K et al. 2000 Statistical pattern recognition: a review *IEEE Trans. Pattern Anal. Mach. Intell.* **22(1)**, 4-37.
- [9] Kwak N and Choi C H 2002 Input feature selection by mutual information based on Parzen window *IEEE Trans. Pattern Anal. Mach. Intell.* **24(12)**, 1667-71.
- [10] Kalyani Mali and Samayita Bhattacharya 2013 Soft computing on medical-data (SCOM) for a countrywide medical system using data mining and cloud computing features *Global Journal of Computer Science and Technology Cloud and Distributed* **13(3)**.

- [11] Ning Wang, Guixia Kan, 2012 A monitoring system for type 2 diabetes mellitus *IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*.
- [12] Li Shuaishuai, Zhang Enke, Li Min and Pan Wei 2015 Research on the effectiveness of application of diabetes management APP. *China Med Dev* **30(08)**.
- [13] Sowjanya K 2015 MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices *IEEE International Advance Computing Conference (IACC)*.
- [14] Gang Shi, Shanshan Liu and Ding Ye 2015 Design and implementation of diabetes risk assessment model based on mobile things *7th International Conference on Information Technology in Medicine and Education*.
- [15] Manjunath A S and Jayaram M A 2011 Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima indians diabetes *International Journal on Soft Computing* **2(2)**.
- [16] Ping Chang and Jeng-Shong Shih 2002 The application of back propagation neural network of multi-channel piezoelectric quartz crystal sensor for mixed organic vapours *Tamkang Journal of Science and Engineering* **5(4)**, 209-217.
- [17] Pradipta Maji and Sankar K Pal 2010 Fuzzy-rough sets for information measures and Selection of relevant genes from microarray data *IEEE transactions on systems, man, and cybernetics—part b: cybernetics* **40(3)**.
- [18] Sellappan Palaniappan and Tan Kim Hong 2008 Discretization of continuous valued dimensions in OLAP data cubes *International Journal of Computer Science and Network Security* **8(11)**.
- [19] Waleed Alomoush, Khairuddin Omar, Ayat Alrosan , Yazan M. Alomari , Dheeb Albashish and Ammar Almomani 2018 Firefly photinus search algorithm *Journal of king saud university-computer and information sciences*.
- [20] Dorigo M, Birattari M, Stutzle T 2006 Ant colony optimization *IEEE Comput. Intell. Mag.* **1**, 28-39.
- [21] Karaboga D, Basturk B 2007 A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm *J. Global Optim.* **39**.
- [22] Kennedy J and Eberhart R 1995 *IEEE International Conference on Neural Networks, Particle swarm optimization* **4**, 1942-48.
- [23] Waleed Alomoush, Khairuddin Omar, Ayat Alrosan , Yazan M. Alomari , Dheeb Albashish and Ammar Almomani 2018 Firefly photinus search algorithm *Journal of king saud university-computer and information sciences*.
- [24] Om prakash varma, Deepti agarwal and Tejna Patodi 2016 Opposition and dimensional based modified firefly algorithm *Expert systems with Applications* **44**, 168-76.
- [25] Hui Wang, Wenjun Wang, Xinyu Zhou, Hui Sun, Jia Zhao and Xiang Yu 2017 Firefly algorithm with neighborhood attraction *Information Science*, 374-87.