International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 1, February 2022, pp. 524~533
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i1.pp524-533

A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus

Ratna Patil[1,2], Sharvari Tamane[3], Shitalkumar Adhar Rawandale[4], Kanishk Patil[5]
[1]Dr Babasaheb Ambedkar Marathwada University, Aurangabad, India
[2]Department of Computer Engineering, Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India
[3]Department of Information Technology, MGM University's Jawaharlal Nehru Engineering College, Aurangabad, Maharashtra, India
[4]Department of Mechanical Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India
[5]Department of Surgery, Grant Govt. Medical College and Sir J. J. Group of Hospital, Mumbai, Maharashtra, India

Article Info

Article history:

Received Apr 10, 2021
Revised Jul 21, 2021
Accepted Aug 8, 2021

Keywords:

Classification
Mayfly algorithm
Optimization
Support vector machine
Type 2 diabetes mellitus

ABSTRACT

Diabetes mellitus is a chronic disease that affects many people in the world badly. Early diagnosis of this disease is of paramount importance as physicians and patients can work towards prevention and mitigation of future complications. Hence, there is a necessity to develop a system that diagnoses type 2 diabetes mellitus (T2DM) at an early stage. Recently, large number of studies have emerged with prediction models to diagnose T2DM. Most importantly, published literature lacks the availability of multi-class studies. Therefore, the primary objective of the study is development of multi-class predictive model by taking advantage of routinely available clinical data in diagnosing T2DM using machine learning algorithms. In this work, modified mayfly-support vector machine is implemented to notice the prediabetic stage accurately. To assess the effectiveness of proposed model, a comparative study was undertaken and was contrasted with T2DM prediction models developed by other researchers from last five years. Proposed model was validated over data collected from local hospitals and the benchmark PIMA dataset available on UCI repository. The study reveals that modified Mayfly-SVM has a considerable edge over metaheuristic optimization algorithms in local as well as global searching capabilities and has attained maximum test accuracy of 94.5% over PIMA.

Corresponding Author:

Ratna Patil
Dr Babasaheb Ambedkar Marathwada University, Aurangabad, India
Department of Computer Engineering, Noida Institute of Engineering and Technology
19, Institutional Area, Knowledge Park II, Greater Noida, Uttar Pradesh 201306, India
Email: ratna.nitin.patil@gmail.com

1. INTRODUCTION

Diabetes mellitus is one among the most widespread, long-lasting diseases plaguing humans worldwide. Approximately 425 million people have been afflicted globally at present and it is predicted that up to 700 million people will be affected by 2045 [1]. It is a metabolic condition caused by raised level of blood glucose leading to health issues such as myocardial infarction, stroke, renal failure, blindness, neuropathy, gangrene and an increased predilection to infections. India has the highest population of diabetes mellitus patients after China in the world [2]. Type 1 (T1DM), type 2 (T2DM) and gestational diabetes (GDM) are the three types of diabetes. In this, type 1 is a condition where the pancreas stopped secreting the insulin so that external injection of it is necessary to maintain the insulin level in the body. In type 2 diabetes, insulin is not utilized properly by the body which needs proper diet, exercises and in some cases they take tablets orally to compensate the insulin level. Type 3 diabetes comprises of high level of blood glucose

during pregnancy and disappear after it. From all these types, T2DM is more prominent among the people. Earlier prediction of this type of diabetes will help the people to get rid of it when proper diet and healthy life style were followed.

If T2DM left untreated it can lead to heart attack, kidney failure, stroke and even blindness. One-third of the population go undetected in early stage [3]. Several data mining methods have been used so far detecting the disease. In today's world, machine learning (ML) gears up many innovations much faster with accurate and reliable outputs. In the recent times, several research works focused on employing ML techniques for the prediction of diabetes mellitus [4], [5]. Various data mining methods such as naïve bayes, decision tree, support vector machine (SVM), neural network (NN), fuzzy logic, ant colony optimization (ACO), genetic algorithm (GA) [6], bayesian network, neuro fuzzy [7], particle swarm optimization (PSO) [8], adaptive neuro-fuzzy inference systems (ANFIS) were found to perform well in terms of accuracy of prediction and error rate [9]. Developing a robust model of prediction is a challenging task. For such problems, the optimization process searches for the best possible solution under given circumstances. And so, several evolutionary and swarm-intelligence based nature-inspired metaheuristic algorithms namely ACO, GA, PSO [8], artificial bee colony (ABC), genetic programming (GP) and differential evolution (DE) have been employed to get the optimal solution [10]. Inspired by such meta-heuristic algorithms, Mayfly optimization algorithm is employed in the proposed work for feature selection, followed by SVM classifier for predicting T2DM.

The main objectives of the proposed work are to implement modified mayfly optimization algorithm for the selection of significant features and to employ SVM classifier for the prediction of T2DM. The study compares the performance of the proposed model of T2DM prediction with that of the existing ones. The processes of nuptial dance performed by male mayflies and random flight of female mayflies boost the exploration and exploitation properties which further helps to escape from local minima. The implementation and analysis of results have shown the superiority of mayfly approach. The rest of the research paper is organized in the following manner: Section 2 presents the review of literature, section 3 includes the description of the proposed methodology, section 4 displays the results obtained for the proposed method, followed by section 5 which concludes the research work.

2. LITERATURE REVIEW

In recent years, people suffering from diabetes have increase manifolds in India and in the world as well. Numerous machine learning approaches as well as data mining techniques like artificial neural network (ANN), SVM, k-nearest neighbors (KNN), decision tree, Extreme learning machine have emerged and have been utilized as an aid in detection of diabetes. Similarly, there has been increasing trend among the researchers on use of metaheuristic optimization algorithms to improve the efficacy of prediction models of diabetes [11]. Combinational (hybrid) approaches improve generalization ability, increase result accuracy and robustness of the model [12].

Zou *et al.* [13] had implemented three classifiers namely neural network, random forest, and decision tree. Authors had carried out comparison of classifiers on Luzhou and PIMA dataset and analyzed it. The study highlights that random forest method was superior to decision tree and neural network methods. The dimensionality of dataset has been reduced with the use of principal component analysis (PCA) and minimum redundancy maximum relevance (MRMR). Based on the experimental analysis, authors had concluded that the accuracy of classification by utilizing PCA was inferior to the accuracy obtained using all the features. But the accuracy obtained using MRMR was better. The outcome however showed that prediction utilizing random forest algorithm could achieve utmost accuracy (ACC) of 80.84% without eliminating any feature on PIMA dataset. This accuracy can be improved by using metaheuristic approaches. Alalwan [14] have used various data mining algorithms like SVM, multilayer perceptron (MLP), naïve bayes, random forest, logistic regression and J.48 and self organizing maps (SOM) to develop predictive model for diabetes on the PIMA dataset. Authors had suggested self-organizing maps to improve upon the accuracy of the prediction. Choi *et al.* [15] have developed prediction model to classify T2DM using electronic medical records. Total number of features extracted were 28 from the electronic medical records for this study. Different machine learning algorithms namely logistic regression, quadratic discriminant analysis, KNN , and linear discriminant analysis, were employed and their performances were compared. Among these algorithms logistic regression had outperformed the other algorithms. Authors have observed that other three algorithms had not shown a statistically significant difference in their results. Kopitar *et al.* [16] compared various machine learning based predictive models, such as, random forest, regularized generalized linear model, extreme gradient boosting, light gradient boosting and commonly employed regression approaches to predict T2DM. However, the results of this study did not show any clinically relevant improvement in the performance. The authors concluded that the highest graded attributes for predicting type 2 diabetes are hyperglycemia history, physical activities, age, high density lipo protein

(HDL), cholesterol, triglycerides, and use of antihypertensive drugs. In this work, authors have suggested use of blending and stacking of different prediction models to improve the prediction accuracy. Singh and Singh [17] proposed stacking-based multi-objective evolutionary ensemble framework for predicting type 2 diabetes mellitus. In this work, PIMA Indian diabetes (PID) has been taken for testing. Different feature selection methods to select informative attributes and filter out the irrelevant attributes and evolutionary multi-objective approaches, individual or hybrid ones can be employed to obtain better Pareto-optimal fronts. Patil *et al.* [18] experimented the cultural algorithms with a series of emerging algorithms with the aim of extracting cultural evolution by following the dual inheritance. The conventional genetic algorithm selects the fittest to converge towards the results indeed does not taken into account the diversity of the population. This literature used cultural algorithms with GA using their operator to enhance the efficiency of the network. Metaheuristic algorithms efficiently explores the search space so as to find good (near-optimal) feasible solutions and does not provide guarantee of local or global optimality. The solution is improved using iterative approach and combines one or more good solutions to generate new solutions. It was observed from the literature study that these smart metaheuristic algorithms have shown high efficacy in this domain so the proposed model using metaheuristic model was implemented.

3. PROPOSED MODIFIED MAYFLY-SVM APPROACH

In the proposed model, raw diabetic dataset is initially subjected to data preprocessing, where the missing cells are filled and unwanted data are removed. The preprocessed data is sent to the modified mayfly optimizer for selecting the features which are significant for the prediction of the disease. Rest of the feature columns are removed. Now the data with selected set of features is sent for data splitting, where the data is split into training set (80%) and testing set (20%). The proposed SVM model is trained using the training set in the training phase. In the testing phase, the fully trained SVM model predicts the existence of T2DM for the given test input sample. The following sub sections explain how the feature selection algorithm and prediction model works.

3.1. Mayfly optimization algorithm for feature selection

Feature selection is a binary optimization problem. Modified mayfly optimization algorithm [19] is applied to choose relevant features and drop the insignificant ones. It can be considered as a modification of PSO [20] and amalgamates major advantages of PSO, GA [21] and FA [19]. Mayfly algorithm which works on the basis of mayflies' behavior, renders an effective hybrid algorithm. When researchers tried to improve PSO algorithm's performance using local search and crossover techniques, they were able to produce good results. This is because, PSO's only demerit is the convergence at local optima, and so modifications have to be made in such a way that optimum point is guaranteed while operating in spaces of high dimensions. Mayfly is inspired from the mayflies' social behavior and their process of mating. It is assumed that, when the mayflies are hatched from the eggs, they are adults and are fit for survival, irrespective of their lifespan. The flowchart of the mayfly optimization algorithm is given in Figure 1. The location of every mayfly in the search dimension is a possible solution to the optimization problem. The working of the algorithm is explained as follows. In the beginning, 2 mayflies (a female from the female population and a male from the male population) are created randomly. The search space created is of d-dimensions and the vector of possible mayfly candidate placed in the search space is x=($x1,\ldots,xd$). The fitness of each candidate is computed using a cost function f(x) which is also known as the objective function. The gathering pattern of the male mayflies infer that the position of one mayfly is changed with respect to the experience of its own and of its adjacent mayflies. Let x_i^t be the current position of ith mayfly in a search dimension of time step t, having a position changed by the sum of the velocity v_i^{t+1} and the position. This is expressed by (1):

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (1)$$

with $x_i^0 \sim U(x_{min}, x_{max})$.

In contrast to the males, the female mayflies avoid swarming. Instead, they fly to the swarm of male mayflies for breeding. Let y_i^t be the current position of i[th] female mayfly in a search dimension of time step t, having a position changed by the sum of the velocity v_i^{t+1} and the position. This is expressed as (2):

$$y_i^{t+1} = y_i^t + v_i^{t+1} \quad (2)$$

with $y_i^0 \sim U(y_{min}, y_{max})$. The velocity v= ($v1,\ldots,vd$) of a mayfly is defined as the change in the position of the mayfly, and the direction towards which the mayfly flies is an interaction which happens dynamically between the 2 individuals and experiences of flying socially.

Mayflies maintain few meters of distance above the surface of the water to perform the nuptial dance. And so, it is assumed that the male mayflies cannot move in high speed and can only move in a constant speed. As a result of this, the velocity of a male mayfly i can be expressed by (3):

$$v_{ij}^{t+1} = v_{ij}^{t} + a_1 e^{-\beta r_p^2 \left(pbest_{ij} - x_{ij}^{t}\right)} + a_2 e^{-\beta r_g^2 \left(gbest_j - x_{ij}^{t}\right)} \quad (3)$$

where v_{ij}^{t} represents the velocity of male mayfly in dimension $j = 1,2, \ldots,$ n at the time period t, x_{ij}^{t} denotes the location of the i^{th} male mayfly in dimension j at time step t, a_1 and a_2 are constants referring to positive attraction applied to scale the contributions of the cognitive component and social component, respectively. In addition to this, $pbest_i$ represents the personal best position of the mayfly i.

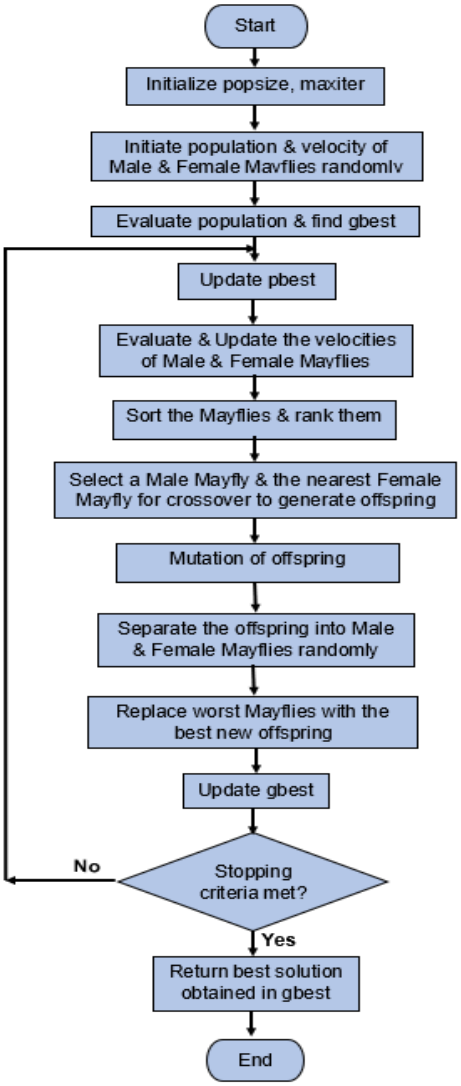


Figure 1. Flowchart of mayfly optimization algorithm

In the proposed work, attraction process of female mayflies is implemented by finding the nearest male mayfly instead of random attraction. This will maximize the convergence behavior of the optimization. Therefore, the velocities of the female mayflies can be evaluated by (4).

$$v_{ij}^{t+1} = \begin{cases} v_{ij}^{t} + a_2 e^{-\beta r_{mf}^2 \left(x_{ij}^{t} - y_{ij}^{t}\right)} & if\ f(y_i) > f(x_i) \\ v_{ij}^{t} + fl \times r & if\ f(y_i) \leq f(x_i) \end{cases} \quad (4)$$

Where v_{ij}^{t} represents the velocity of female mayfly in dimension $j = 1,2, \ldots, \mathrm{n}$ at the time period t, y_{ij}^{t} denotes the location of the i^{th} female mayfly in dimension j at time step t, a_2 is the constant referring to positive attraction, and β indicates a fixed coefficient of visibility that is used to bound the visibility of the mayfly to other mayflies, while r_{mf} is the representation of Cartesian distance between female and male mayflies. The coefficient of random walk is denoted by fl and computed as shown in (5).

$$fl_{iter} = fl_0 * \delta^{iter} \tag{5}$$

This is applied when the female is not attracted to the male mayfly. This makes the female mayfly to fly randomly, and the random value falls in the interval [-1, 1], denoted by r and the current iteration number is iter and δ lies in the interval between [0, 1].

The personal best position $pbest_{ij}$, at the next time step t+1, can be computed by (6):

$$pbest_i = \begin{cases} x_i^{t+1} & if\ f(x_i^{t+1}) < f(pbest_i) \\ remains\ the\ same & otherwise \end{cases} \tag{6}$$

where $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ denotes the cost function, that evaluates the solution quality. The global best position $gbest$ at time step t, can be expressed by (7):

$$gbest \in \{pbest_1, \ldots, pbest_N | f(cbest)\} = \min\{f(pbest_1), \ldots, f(pbest_N)\} \tag{7}$$

where N denotes the total count of male mayflies, β indicates a fixed coefficient of visibility that is used to bound the visibility of the mayfly to other mayflies, while r_p is the representation of Cartesian distance between x_i and $pbest_i$ and r_g is the representation of Cartesian distance between x_i and $gbest$. The values of the distances can be calculated using (8):

$$\|x_i - X_i\| = \sqrt{\sum_{j=1}^{n} (x_{ij} - X_{ij})^2} \tag{8}$$

where x_{ij} represents the jth element of mayfly i and X_i corresponds to $pbest_i$ or $gbest$.

Particularly, all the mayflies adjust their trajectories toward their personal best position ($pbest$) achieved so far, and the global best position in the entire swarm achieved so far ($gbest$). A stochastic element to the algorithm is primarily given by best mayflies at a particular time by performing the nuptial dance. This dance is mathematically represented by (9):

$$v_{ij}^{t+1} = g * \ v_{ij}^{t} + d * r \tag{9}$$

where, g is the inertia weight, nuptial dance is represented by d and damping ratio r is a random value falls in the interval [-1, 1]. The nuptial dance coefficient gradually decreases and computed by (10).

$$d_{iter} = d_0 * \delta^{iter} \tag{10}$$

Nuptial dance coefficient's initial value is denoted as d_0 and the current iteration number is iter and δ lies in the interval between [0,1].

The crossover operation is carried out by selecting male mayfly and then nearest female mayfly and offspring are yielded using (11) and (12). Equation (13) is used for selecting the nearest female mayfly.

$$first\_offspring = \ rnd * male + (1 - rnd) * female \tag{11}$$

$$second\_offspring = \ rnd * female + (1 - rnd) * male \tag{12}$$

$$Distance_{ij} = abs(\ mean(Position_i) - mean(\ (Position_j)\ ) \tag{13}$$

The offspring generated are mutated with use of random resetting to enhance model's exploration ability. In the proposed model for feature selection, the mayfly algorithm is selected and hybridized as its

convergence behaviour is exceptional. Mayfly algorithm is opted for feature selection because nuptial dance (performed by male mayfly) and random walk (by female mayfly) enhances the balance between exploration of search space and its exploitation and helps to escape from local optimum. The individual encoding and fitness functions are two important steps to be determined before discussing about the details of metaheuristic approach for feature selection. Index based encoding is used for individual representation. An individual consists of m genes an attribute is considered selected if its corresponding index occurs in at least one of the genes of the individual. In this wrapper based model, rastrigin function is used for the fitness function. In this work, attraction process of female mayflies is implemented by finding the nearest male mayfly instead of random attraction. The modified mayfly algorithm is implemented for selecting important features and then SVM is employed for prediction of T2DM. The convergence behavior of the mayfly method is also exceptional as it reaches the best overall solution in the early iterations most of the times [22]. Pseudocode for the proposed modified mayfly algorithm is mentioned below.

Mayfly algorithm

```
Input: population_size, max_no_of_iterations
Output: return the best solution obtained x=(x1, x2, …, xd )
Random initiation of population and velocity of male mayflies
Initiate population and velocity of female mayflies randomly
Evaluate solutions
Find the global best and assign it to gbest using the (7)
For i=1 to max_no_of_iterations
For j=1 to population_size
Evaluate and update male and female mayflies' velocities using the (3) and 4 respectively
Update the pbest by using (6)
End for
Perform crossover/mating of male mayfly with the nearest female mayfly instead of random
attraction to generate offspring using (11) and (12). The (13) is used to calculate the
nearest female
Mutation is applied to generate offspring
Separate offspring into male and female mayflies randomly
Substitute to worst mayflies with the best new solutions
Update the gbest by the (7)
End for
```

SVM has origins in statistical learning theory. As a task of classification, it looks up for optimal decision boundary (hyperplane) separating the tuples of one class from another. SVM is the quintessential supervised classifier that maximizes the margin to maximize generalization and reduces overfitting and underfitting by using various kernel functions that aid nonlinear separation [5]. They are more effective in high dimensional spaces. Hence, support vector machine with rbf kernel (rbf-SVM) is utilized for the prediction of T2DM in this study. The features that are chosen using Mayfly algorithm are sent to SVM for prediction.

4. RESULTS AND DISCUSSION

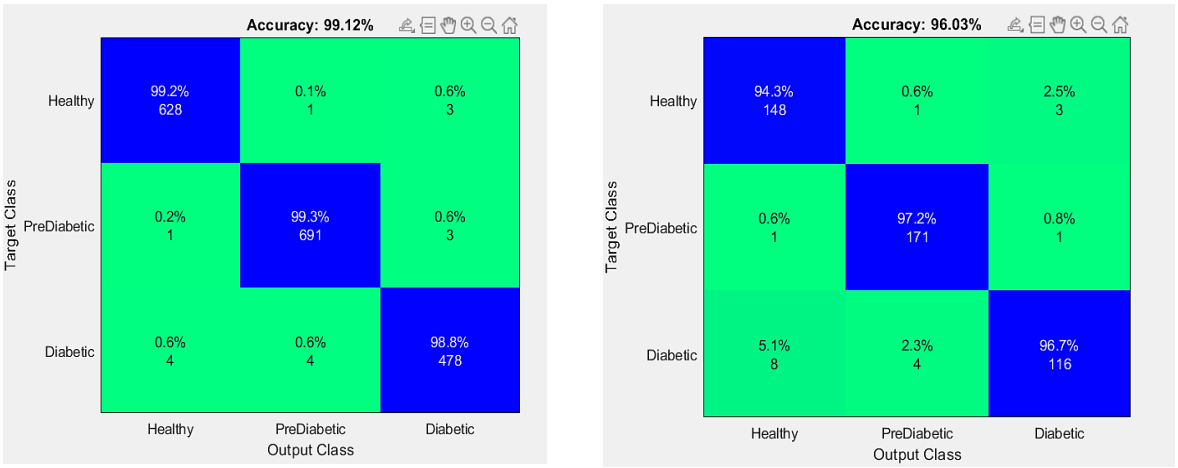
4.1. Real time dataset

Realtime dataset has been gathered from 1133 patients admitted in local hospitals with due consent. This dataset has a substantial advantage over the PIMA dataset as it eliminates the selection bias present in the latter. PIMA dataset contains records of only females equal to or above 21 years of age. Another limitation of the PIMA dataset is that it only has 2 class labels - healthy and diabetic whereas in this dataset, a pre-diabetic label has also been introduced. Lifestyle modifications in pre diabetics will halt the disease progression. Total number of features collected from each person was 33. They are: gender, height, age, weight, waist circumference, body mass index (BMI), systolic and diastolic blood pressure, HbA1c level, HDL cholesterol, low density lipoprotein (LDL) cholesterol, very low density lipoproteins (VLDL), serum creatinine (impaired kidney function or disease), Triglyceride level, fasting blood glucose, post prandial glucose, family history, medications for high blood pressure, Physical activity/exercise (minimum 30 minutes daily), vegetable/fruit intake daily, drinking, smoking status, excess hunger, excess thirst, frequent urination, itchy skin, increased fatigue, depression and stress, frequent infection, poor wound healing and blurred vision.

4.2. Simulation results

In this study, 25 runs of the developed modified mayfly-SVM were performed. Synthetic minority oversampling technique (SMOTE) is employed for increasing the number of cases in the dataset in a balanced way. A comprehensive analysis of the obtained results was conducted. The accuracy, error rate, sensitivity and specificity of both approaches were evaluated and compared with other approaches. Mayfly

optimization combines the advantages of swarm intelligence and evolutionary algorithms. Significant features are selected and sent to fine Gaussian SVM classifier model for prediction. The confusion matrices of Training set and testing set are given in Figure 2 over the collected dataset from local hospitals. The same model of prediction is tested with benchmark PIMA dataset available on UCI repository [23]. The SMOTE method has increased the instances of real-time and PIMA dataset in a balanced way. The increased number of instances can be seen in the confusion matrices.



(a) (b)

Figure 2. Confusion matrix; (a) training set and (b) testing set over real time dataset

Performance evaluation of the developed model based on parameters such as accuracy and error rate only are not adequate, hence statistical measures have also been computed to describe the models. The average model performance of modified mayfly-SVM is presented in Table 1. The indicators of prediction accuracy like sensitivity, specificity, F1-score, Mathew's correlation coefficient and kappa coefficient are above 0.9 and approaching 1 which can be interpreted as better performing prediction model.

Table 1. Performance of modified mayfly-SVM model

Performance Measure	PIMA	Realtime
Testing accuracy %	94.46	96.03
Error	0.06	0.04
Sensitivity	0.98	0.96
Specificity	0.88	0.98
Precision	0.94	0.96
False positive rate (FPR)	0.12	0.02
F1_score	0.96	0.96
Matthews Correlation Coefficient (MCC)	0.87	0.94
Kappa Coefficient	0.87	0.91

Performance of modified mayfly-SVM model on PIMA and real-time dataset was assessed and presented in Figure 3. It is observed that the results are improved further by the proposed model. After several runs of the experimental study, the selected features varied but the maximum number of times the following features were selected. weight, waist circumference, BMI, Hba1c level, LDL, HDL cholesterol, VLDL, serum creatinine, triglyceride level, fasting blood glucose, post prandial glucose, family history, excess hunger, excess thirst, frequent urination, infection, poor wound healing.

4.3. Performance comparison

The performances of the proposed models are compared with the existing approaches using PIMA benchmark dataset [22]. This data is from National Institute of Diabetes and Digestive and Kidney Diseases. The main goal is to anticipate whether the patient is diabetic or not, based on the data collected for diagnosis.

In PIMA dataset, all the patients are females and are equal to or above 21 years of age. The features present in the dataset are pregnancy count, concentration of plasma glucose (obtained from a 2-hour oral glucose tolerance test), diastolic blood pressure (measured in mm Hg), triceps skin fold thickness (measured in mm), 2 hours of serum insulin (measured in mu U/ml), body mass index (weight in kg/(height in m^2), pedigree function of diabetes, age (in years) and one class variable (class 0 for healthy and class 1 for diabetic). It's advisable to analyze the results of proposed methods with the work done by other researchers since last 5 years using available performance metrics. Selected approaches were systematically studied, analyzed and compared with the proposed models based on the prediction accuracy and depicted in Table 2.

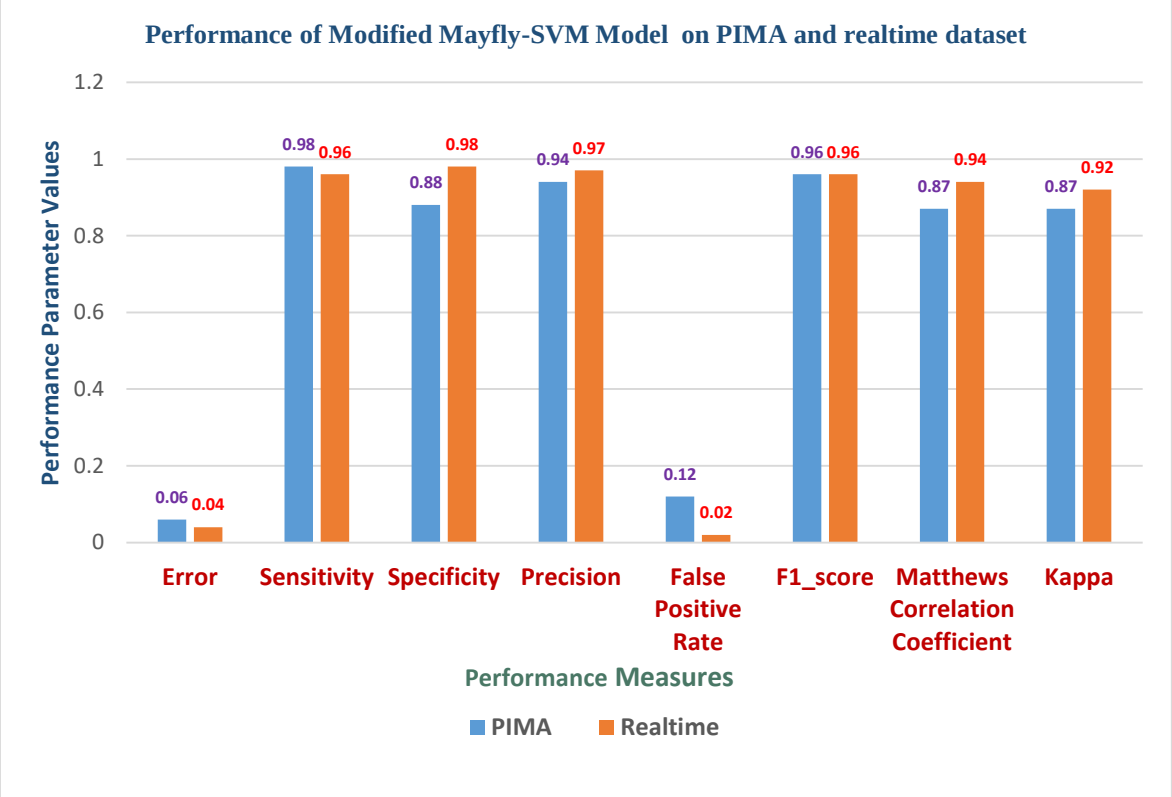


Figure 3. Performance of modified mayfly-SVM model on PIMA and realtime dataset

Table 2. Results of existing type 2 diabetes predictive models in the literature

Sr No	Techniques	Ref No.	Results (Accuracy %)
1	Back propagation neural network	[24]	81
	Decision tree		73.82
2	SVM	[25]	65.1
	Naïve Bayes		76.30
3	SW-FFANN	[26]	91.66%
4	ANN	[27]	87.3
5	DNN	[28]	77.8
	SVM		77.6
6	RF and Gradient boosting classifiers	[29]	90
7	Auto MLP	[30]	88.7
8	Stacking based multi objective evolutionary ensemble	[17]	83.8
9	PCA+ANN	[31]	75.7
10	PCA+kmeans+LR	[32]	79.94
11	PCA and minimum redundancy maximum relevance	[13]	77.21
12	Self-organizing map	[14]	84
13	RMLP (Resampling version of MLP)	[12]	79.30
14	Gaussian Fuzzy decision tree	[33]	75
15	PSO-ANN	[8]	80%
16	Cultural algorithm+ANN	[18]	79%
17	Genetic algorithm with multi objective evolutionary Fuzzy classifier	[34]	83.04
18	Genetic algorithm+kNN	[21]	82.3
19	Modified mayfly SVM (current work)	Current Work	94.5

The comparison of the obtained predictive accuracy of current work with the existing techniques over PIMA has been depicted at Figure 4. The graph indicates that the current work has outperformed existing methods in the literature. The maximum predictive accuracy using the modified Mayfly-SVM was 94.5% and has outperformed the existing models. There are three techniques which have shown prediction accuracy more than 90%. These three methods are modified mayfly-SVM of current work, small-world feed forward artificial neural network (SW-FFANN) [26] and RF and gradient boosting classifiers [29]; the first method is one of the models developed by me and others are existing one which have potential to improve further for achieving even better accuracy in future.

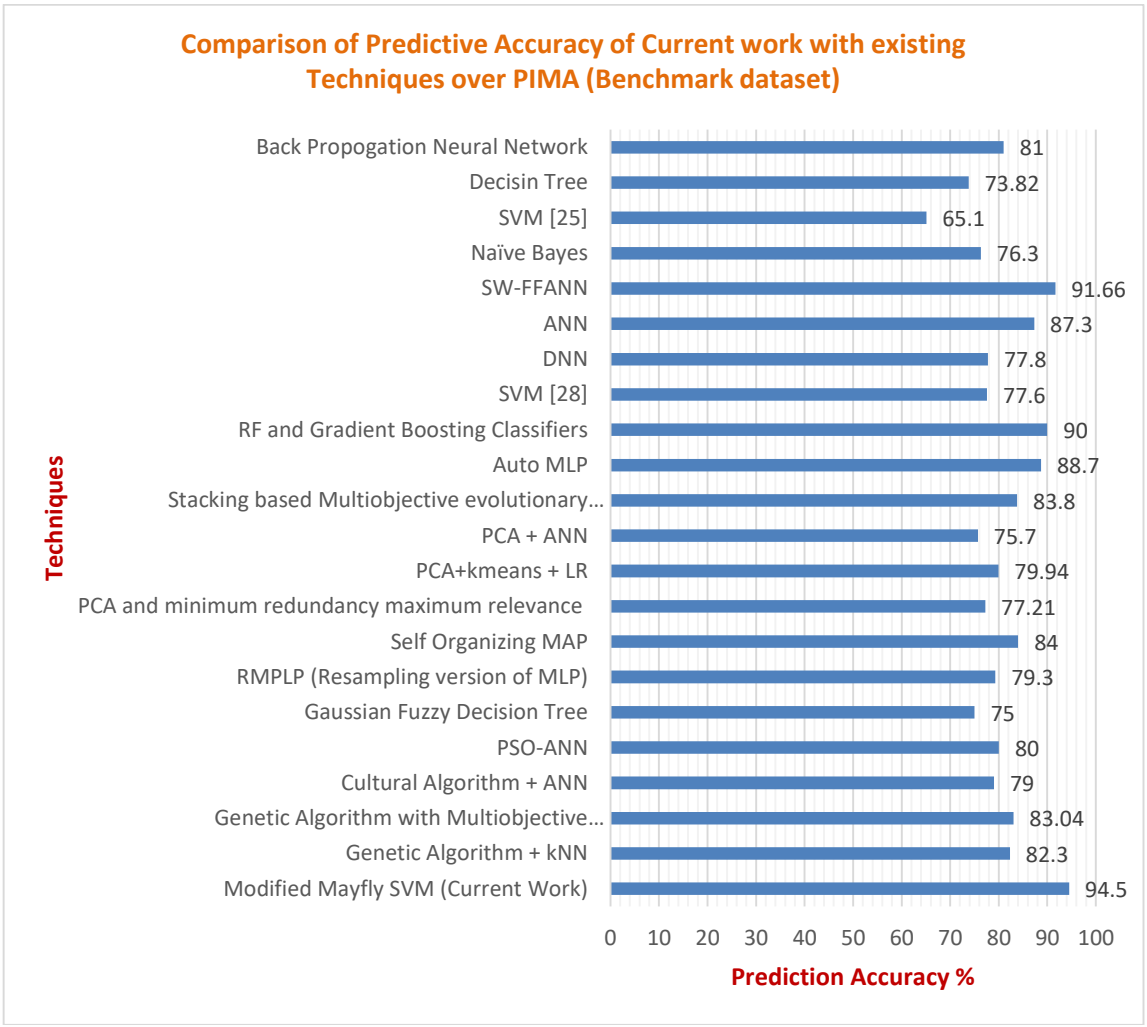


Figure 4. Comparison of obtained predictive accuracy of current work with existing techniques over PIMA dataset

5. CONCLUSION

The diagnosis of T2DM has profound implications on an individual from medical as well as financial perspective. The proposed models can diagnose this disease early (prediabetic stage), so that the physician and patient can work towards prevention and mitigation of complications caused by T2DM. As an effort to improve the performance of the prediction of diabetes mellitus, an effective framework based on mayfly-SVM approach was proposed. In this work, mayfly optimization algorithm was employed for feature selection and SVM was developed for predicting T2DM. Extensive experimentation on real-world and benchmark data set for validating the efficiency of proposed model to detect T2DM at the earlier stage and monitoring the disease at several stages was done and the accuracy of the system is compared with all other systems and proved to be efficient from the results obtained. As a future work reasonable and rational health suggestions can be provided to the high-risk group.

REFERENCES

[1] International Diabetes Federation. “IDF Diabetes Atlas.” 9th Ed. diabetesatlas.org 2019. https://www.diabetesatlas.org/ (accessed Sep. 2, 2021).

[2] National Health Portal of India. “diabetes-mellitus.” nhp.gov.in. https://www.nhp.gov.in/disease/digestive/pancreas/diabetes-mellitus (accessed Sep. 2, 2021).

[3] American Diabetes Association, “Classification and diagnosis of diabetes: standards of medical care in diabetes-2018,” *Diabetes care 41.Supplement 1*, 2018, doi: 10.2337/dc18-S002.

[4] R. N. Patil and S. C. Tamane, "A survey paper on evolving techniques for the prediction of type 2 diabetes," *International Journal of Computer Science and Information Security*, vol. 4, no. 10, pp. 329-333, 2016.

[5] T. Mitchell, *Machine Learning*, McGraw Hill, 2017.

[6] R. Patil, S. Tamane and N. Rawandale, "Hybrid ANFIS-GA and ANFIS-PSO based models for prediction of type 2 diabetes mellitus," in *Computational Methods and Data Engineering*, vol. 1227, pp. 11-23, 2021, doi: 10.1007/978-981-15-6876-3_2.

[7] R. Patil, S. Tamane and K. Patil, "Self organising fuzzy logic classifier for predicting type-2 diabetes mellitus using ACO-ANN," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 7, pp. 348-353, Jan. 2020, doi: 10.14569/IJACSA.2020.0110746.

[8] R. Patil and S. C. Tamane, "PSO-ANN-based computer-aided diagnosis and classification of diabetes," in *Smart Trends in Computing and Communications*, vol. 165, pp. 11-20, 2020, doi: 10.1007/978-981-15-0077-0_2.

[9] H. Lai, H. Huang, K. Keshavjee, A. Guergachi and G. X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, 2019, doi: 10.1186/s12902-019-0436-6.

[10] X.-S. Yang, *Nature-inspired metaheuristic algorithms*, Luniver Press, 2010.

[11] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, 2018, doi: 10.1016/j.aci.2018.12.004.

[12] M. R. Nalluri, K. Kannan, M. Manisha, and D. S, Roy, "Hybrid disease diagnosis using multiobjective optimization with evolutionary parameter optimization," *Journal of Healthcare Engineering*, vol. 2017, 2017, Art. no. 5907264, doi: 10.1155/2017/5907264.

[13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, no. 9, pp. 1-10, 2018, doi: 10.3389/fgene.2018.00515.

[14] S. A. D. Alalwan, "Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 1, pp. 88-95, 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95.

[15] B. G. Choi, S. W. Rha, S. W. Kim, J. H. Kang, J. Y. Park and Y. K. Noh, "Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks," *Yonsei Medical Journal*, vol. 60, no. 2, pp. 191-199, 2019, doi: 10.3349/ymj.2019.60.2.191.

[16] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, 2020, Art. no. 11981, doi: 10.1038/s41598-020-68771-z.

[17] N. Singh and P. Singh, "Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 1-22, 2019, doi: 10.1016/j.bbe.2019.10.001.

[18] R. Patil, S. Tamane and K. Patil, "An experimental approach toward type 2 diabetes diagnosis using cultural algorithm," in *ICT Systems and Sustainability*, vol. 1270, pp. 405-415, 2021, doi: 10.1007/978-981-15-8289-9_39.

[19] K. Zervoudakis and S. Tsafarakis, "A mayfly optimization algorithm," *Computers and Industrial Engineering*, vol. 145, 2020, Art. no. 106559, doi: 10.1016/j.cie.2020.106559.

[20] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, pp. 1942-1948, vol. 4, doi: 10.1109/ICNN.1995.488968.

[21] R. N. Patil and S. Tamane, "Upgrading the performance of KNN and naïve bayes in diabetes detection with genetic algorithm for feature selection," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, pp. 2456-3307, 2018.

[22] T. Bhattacharyya, B. Chatterjee, P. K. Singh, J. H. Yoon, Z. W. Geem and R. Sarkar, "Mayfly in harmony: a new hybrid meta-heuristic feature selection algorithm," in *IEEE Access*, vol. 8, pp. 195929-195945, 2020, doi: 10.1109/ACCESS.2020.3031718.

[23] “UCI Repository." kaggle.com. https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed Sept. 2, 2018).

[24] S. Joshi and M. Borse, "Detection and prediction of diabetes mellitus using back-propagation neural network," *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, 2016, pp. 110-113, doi: 10.1109/ICMETE.2016.11.

[25] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578-1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[26] O. Erkaymaz and M. Ozer, "Impact of small-world network topology on the conventional artificial neural network for the diagnosis of diabetes," *Chaos, Solitons and Fractals*, vol. 83, pp. 178-185, 2016, doi: 10.1016/j.chaos.2015.11.029.

[27] N. S. El-Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," *International Journal of Advanced Science and Technology*, vol. 124, pp. 1-10, 2018, doi: 10.14257/ijast.2018.124.0.

[28] S. Wei, X. Zhao and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, 2018, pp. 291-295, doi: 10.1109/WF-IoT.2018.8355130.

[29] S. K. Das, A. K. Mishra and P. Roy, "Automatic diabetes prediction using tree-based ensemble learners," *Journal of Computational Intelligence and IoT*, vol. 2, no. 2, pp. 485-490, 2019.

[30] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," *2017 Intelligent Systems Conference (IntelliSys)*, 2017, pp. 722-728, doi: 10.1109/IntelliSys.2017.8324209.

[31] T. M. Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, no. 16, pp. 1-6, 2019, Art. no. 100204, doi: 10.1016/j.imu.2019.100204.

[32] C. Zhu, C. U. Idemudia and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and k-means techniques," *Informatics in Medicine Unlocked*, vol. 17, pp. 1-7, 2019, Art. no. 100179, doi: 10.1016/j.imu.2019.100179.

[33] K. V. S. R. P. Varma, A. A. Rao, T. S. M. Lakhsmi, and P. V. N. Rao, "A computational intelligence approach for a better diagnosis of diabetic patients," *Computers and Electrical Engineering*, vol. 40, no. 5, pp. 1758-1765, 2014, doi: 10.1016/j.compeleceng.2013.07.003.

[34] V. Ravindranath, S. Ra, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm-based feature selection and MOE fuzzy classification algorithm on pima Indians diabetes dataset," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, 2017, doi: 10.1109/ICCNI.2017.8123815.