

An efficient stacking based NSGA-II approach for predicting type 2 diabetes

Ratna Nitin Patil¹, Shitalkumar Rawandale², Nirmalkumar Rawandale³, Ujjwala Rawandale⁴, Shrishti Patil⁵

¹Department of Artificial Intelligence and Machine Learning, Noida Institute of Engineering and Technology, Greater Noida, India

²Department of Mechanical Engineering, Pimpri Chinchwad College of Engineering, Pune, India

³Department of Medicine, Shri Bhausaheb Hire Government Medical College, Dhule, India

⁴Department of Electronics and Telecommunication Engineering, MIT World Peace University, Pune, India

⁵Seth Gordhandas Sunderdas Medical College, Mumbai, India

Article Info

Article history:

Received Mar 18, 2022

Revised Sep 25, 2022

Accepted Oct 7, 2022

Keywords:

Ensemble

K-nearest neighbors

NSGA-II

Stacking based ensemble

ABSTRACT

Diabetes has been acknowledged as a well-known risk factor for renal and cardiovascular disorders, cardiac stroke and leads to a lot of morbidity in the society. Reducing the disease prevalence in the community will provide substantial benefits to the community and lessen the burden on the public health care system. So far, to detect the disease innumerable data mining approaches have been used. These days, incorporation of machine learning is conducive for the construction of a faster, accurate and reliable model. Several methods based on ensemble classifiers are being used by researchers for the prediction of diabetes. The proposed framework of prediction of diabetes mellitus employs an approach called stacking based ensemble using non-dominated sorting genetic algorithm (NSGA-II) scheme. The primary objective of the work is to develop a more accurate prediction model that reduces the lead time i.e., the time between the onset of diabetes and clinical diagnosis. Proposed NSGA-II stacking approach has been compared with Boosting, Bagging, Random Forest and Random Subspace method. The performance of Stacking approach has eclipsed the other conventional ensemble methods. It has been noted that k-nearest neighbors (KNN) gives a better performance over decision tree as a stacking combiner.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ratna Nitin Patil

Department of Artificial Intelligence and Machine Learning, Noida Institute of Engineering and Technology
Greater Noida, Gautam Budh Nagar, Uttar Pradesh, India

Email: ratna.nitin.patil@gmail.com

1. INTRODUCTION

Diabetes mellitus is composed of a group of metabolic disease predominantly characterized by high blood glucose levels. Type 2 diabetes mellitus (T2DM) is amongst the most common chronic disorders plaguing the modern life, often complicating other pre-existing diseases. It is caused by varying levels of insulin resistance, decreased insulin secretion, and increased production of hepatic glucose by the liver. A vast majority of diabetes cases in the world are due to type 2 DM. The diagnosis of DM has significant ramifications to an individual's health as well as financial status.

Early diagnosis of type 2 diabetes is still a formidable task at large for the medical service sector. Development of newer innovative models will be required for timely detection of diabetes. This in turn will reduce the dreaded complications and also the burden on the health care system. In today's high-tech world and evolution of technology, techniques like machine-learning have boundless role to play in predicting

T2DM [1]. They have become increasingly popular because of their cost-effectiveness, robustness, generalizability and universal applicability. Algorithms like k-nearest neighbor (KNN) classifier, Bayes theorem [2], support vector machines (SVM), linear discriminant analysis, decision trees, random forest, fuzzy-based methods and ensemble learning classifiers have been used [3]–[5].

Recently, researchers have developed prediction models to diagnose T2DM. To my knowledge, no study, was however aimed at predicting an appropriate diagnosis code especially using the ensemble methods for T2DM patients. Most of all published work (literature) lacks the availability of multi-class studies. On that account, the main aim is to build multi-class predictive model by using available clinical data in analyzing T2DM based on machine learning (ML) techniques such as parallel ensemble technique-bagging and sequential ensemble technique-boosting.

In this work, a NSGA-II stacking model has been implemented to spot the prediabetic stage. For evaluating the performance of suggested model, a comparative study was initiated and was contrasted with algorithms of data mining like decision tree-J48, naïve Bayes (NB), multilayer perceptron (MLP), SVM, and KNN. Proposed models and the data mining algorithms are validated with real clinical data collated from hospitals and the Pima dataset available on University of California Irvine (UCI) machine learning. Moreover, the developed models' performances were computed and analyzed by comparing it with the related work done using hybrid approaches to predict T2DM by other researchers.

2. RELATED WORK

There have been several systems proposed or in vogue to diagnose type 2 diabetes mellitus, however, the accuracy of systems using different techniques of data mining and ML algorithms has not been very high. Of late, researchers tried persistently to increase the prediction accuracy of the systems developed, but there have been no productive results. The systems developed so far have encountered certain issues like: i) use of physiological factors only for diagnosis, ii) testing on a small number of instances, iii) data limited to females only not encompassing males, iv) no multiclass study to facilitate timely detection, and v) no detection of prediabetic stage to improve prognosis treatment

Therefore, with the aim of improving prediction accuracy, a need has arisen to develop a diagnostic system capable of considering all the above-mentioned aspects into account. The model would be multi-class prediction which uses minimal number of assumptions and processing features. Recently, number of people suffering from diabetes increased manifold in India as well as in the world. Diabetes has become a common disease leading to growing interest of researchers in optimization of predictive model for early detection. Several machine learning approaches and data mining techniques like artificial neural network (ANN), decision tree, KNN, SVM, extreme learning machine have become apparent and are being applied in aiding the prediction of T2DM detection of diabetes [6]–[8]. Aseri *et al.* [9] have used metaheuristic algorithms for classifying the covid-19 dataset available on Kaggle. Also, there has been rising trend on use of metaheuristic optimization approaches among the researchers to enhance the effectiveness of prediction models of diabetes. Singh and Singh [10] proposed multi objective optimization technique for minimizing the ensemble complexity and improving the accuracy for classification. The authors have pointed out that prediction accuracy can be improved by using latest and advanced machine learning algorithms and different feature selection methods can be employed for filtering out irrelevant attributes and selecting informative attributes. Diwan [11] had carried out a detailed study in his paper highlighting several methods for predicting T2DM.

In the past authors have used numerous data mining algorithms to implement predictive model for T2DM on the Pima dataset. Lai *et al.* [3] developed prediction models by using gradient boosting machine (GBM) and logistic regression algorithms. The performance of GBM and logistic regression models have been found superior as compared to the performance of random forest (RF) and decision tree models. The performance of models was evaluated using receiver operating characteristic (ROC). The suggested GBM model had ROC of 84.7% with 71.6% sensitivity whereas the logistic regression model had ROC of 84.0% with 73.4% sensitivity. Various researchers have implemented metaheuristic algorithms for severity of disease prediction. Raghavendra and Kumar [12] have experimented on Pima using random forest. The authors have obtained better accuracy by dropping only one feature. The dropping ration can be improved further. Rajni and Amandeep [2] have implemented RB-Bayes technique and obtained 72.9% accuracy over Pima. Sofiana and Sutikno [13] have suggested optimized backpropagation algorithm and observed that training phase improves 12.4 times more than standard backpropagation.

3. MATERIALS

Two datasets Pima and the collected dataset were used in this experiment. Pima dataset which is publicly available has been used by other researchers has 8 attributes and one class label [14]. The data of

1,133 patients admitted in local hospitals with their due consent has been gathered and collated in the second dataset. The collated data has a significant gain over the Pima (benchmark data) as it removes the selection bias present in the latter. The biases are tuples of only females more than twenty-one years of age and it has only two labels-diabetic and healthy. In the collated data, a pre-diabetic class has also been introduced. Lifestyle modifications in pre diabetics will curb the disease (disorder) progression. Total count of attributes gathered from every individual was thirty-three. The features are: gender, weight, height, age, waist (i.e. circumference), body mass index (BMI), systolic and diastolic blood pressure (BP), hemoglobin A1c (HbA1c) level, high-density lipoprotein (HDL) and low-density lipoprotein (LDL) cholesterol, very-low-density lipoprotein (VLDL), serum creatinine, triglyceride level, fasting blood glucose (FBG), post prandial glucose (PPG), family history, medications for high BP, physical activity/exercise (minimum thirty minutes daily), vegetables and fruit eating every day, excess appetite, smoking status, drinking, excess thirst, frequent urination, increased fatigue, itchy skin, frequent infection, depression and stress, poor wound healing and hazy vision.

4. PROPOSED METHOD

Proposed methods of important feature selection and prediction model are discussed below. The flow diagram of suggested NSGA-II stacking model of T2DM prediction is shown in Figure 1. NSGA-II is one of the evolutionary algorithms developed by Deb *et al.* [15] and can be employed to get multiple Pareto optimal solutions in one run.

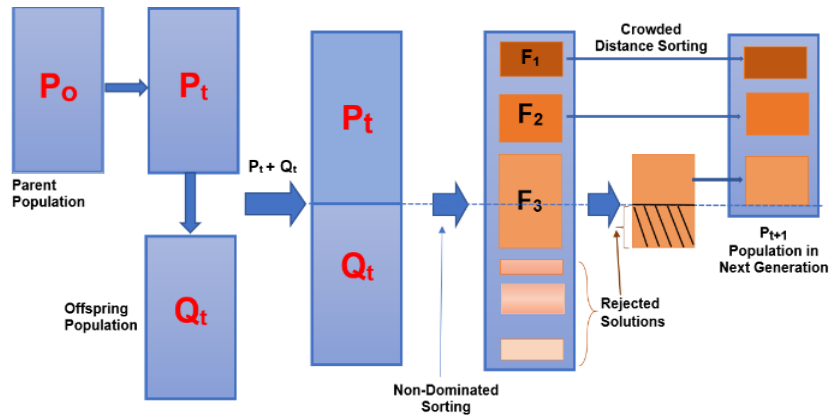


Figure 1. Working of NSGA-II (attributes selection)

4.1. Feature selection

Binary chromosomes were employed to represent the features in this method (1-feature selected, 0-feature not selected). NSGA-II is utilized to pick the important features and rule out the irrelevant ones. It is a multi-objective optimization technique that can be developed to choose minimal number of features and maximize the accuracy. In Figure 1 the process of attribute selection is depicted. This algorithm was found to have high efficacy in the literature review. This technique emphasizes non-dominated sorting. It is notable that elitism present in the algorithm does not permit an already converged Pareto optimum solution to be removed. Population set is initialized randomly and is sorted on the basis of non-dominance. A population set consisting of offspring is created by using binary tournament, mutation and recombination on parent population. A joint population set is created from the population of parents and offspring. The sorting of population is carried out on the basis of non-dominant relations. A new population set is created by including the solutions of the initial front and the subsequent fronts till the population limit is reached to preserve elitism (ensuring that previously obtained best solutions are stored). Every individual is assigned a rank based on their fitness value or on the basis of the front to which the individual belongs to. The fitness value of the first front individuals is assigned 1 and the ones in the second front are assigned as 2, similarly for 3, 4, and till the last front.

Diversity is preserved with the help of crowding distance comparison. The parameter known as crowding distance is computed for all the individuals using (1):

$$n_d = n_d + \frac{f_i(k+1) - f_i(k-1)}{f_i^{max} - f_i^{min}} \quad (1)$$

where, f_i^{min} is the minimum value of i th objective function, f_i^{max} is the maximum value of i th objective function for normalization, i represents the i th objective function value. n_d is initialized to 0.

This crowding distance is defined as the measure of the closeness of one individual with others in the population. If the average crowding distance is large, then the result will be better in terms of population diversity. For the purpose of computing the crowding distance of an individual, the average of all the crowding distances is computed with respect to the neighboring individuals who are in the same front in all the objective test functions (or dimensions) using (1). With the help of the selection of binary tournament, the parents are chosen from the population on the basis of the crowding distance and the rank they hold. For the choice of nondominant solutions, the following partial order methods in (2) is adopted:

$$i < j, \text{ if } \begin{cases} i_{rank} < j_{rank} \\ \text{or} \\ i_{rank} = j_{rank} \text{ and } n_d^i > n_d^j \end{cases} \quad (2)$$

when rank based selection is carried out, the rank of the selected individual is smaller than the others. When two solutions belong to the same Pareto front, the solution with large crowding distance will be selected. In the crossover and mutation processes, offspring are produced from the population selected. The sorting process is again repeated for the current offspring and current population on the basis of non-dominance. Only the P best individuals are chosen, where P indicates the size of the population. Thus, process of selection is carried out on the basis of crowding distance and rank. The pseudocode of the NSGA-II feature selection method is presented as follows.

Pseudocode

Input: Maximum number of generations, probability of crossover, probability of mutation.

Output: Evolved generations based on maximum accuracy and minimum number of features.

1. Initialize population P_t
2. Evaluate objective values of P_t
3. Use Pareto dominance sort to rank individuals
4. Repeat steps 4 - 8 until the stopping criteria is met
 - For $I = 1$ to round $[(P_c * N)/2]$
 - Select (Tournament Selection) two individuals X_1 and X_2 .
 - Perform Crossover (random selection of one point or two point) to obtain offspring population Q .
 - End for
 - For $I = 1$ to round $[(P_m * N)/2]$
 - Select an individual X .
 - Perform mutation to obtain offspring population Q .
 - End for
5. Combine offspring Q and parents P .
6. Use Pareto dominance sort to rank individuals.
7. Compute the crowding distance of individuals in every front by the (1).
8. Select the best N individuals based on calculated ranks using (2) and maximum crowding distance to form the next generation.
9. Extract the best Pareto fronts to present results.

4.2. Prediction model

Ensemble stacking approach (SA) is a technique that puts together different prediction approaches onto a single framework, working at levels or layers. This method presents the meta-learning concept and intends to reduce the generalization errors by decreasing the bias of the generalizers. Initially, training data is used to train base learner models. In the next stage, a meta-algorithm or a combiner is trained for making the final prediction based on the outputs obtained from the base learner models. Such kind of stacked ensembles intend to perform better than any individual base learner models. The process is described as given below:

Step 1: Learn first level classifiers based on the original training dataset.

Step 2: Creation of new dataset with reference to the outputs obtained from the base learners. The predicted outputs obtained from the 1st level learners are called the new features and the actual outputs are set as classes to form the new dataset.

Step 3: Train a 2nd level learner on the basis of dataset newly created. Finalize the training algorithm to be employed for training the second level learner based on the accuracy.

The flow diagram of the suggested model of T2DM prediction is depicted in Figure 2. The dataset is read and is first pre-processed to fill the missing places. Once the data is pre-processed, the parameters of NSGA-II are initialized. The NSGA-II produces Pareto optimal solution set (significant features). The dataset of selected features is split into training set-60%, validation set-20% and testing set-20%. The training set is utilized to construct base learners. Following diverse base learners-linear SVM, radial basis function (RBF)

SVM, Gaussian RBF kernel, KNN-1, KNN-3, KNN-5 and decision tree are harnessed to increase the performance. The predictions on validation set along with the actual labels form the level 1 data which is subsequently used for training the meta learner. Experimentation with four types of meta learners namely bagging, decision tree, linear SVM and KNN was carried out. The base learners were combined with meta learner in identifying the best Meta learner based on the results obtained. It was noted that KNN performed relatively better and has shown encouraging results with respect to the obtained accuracy. Finally, KNN was fixed as a stacking aggregator. KNN classifies records based on the closest training samples in the feature space. Majority voting technique is used to classify the instance amongst its k-nearest neighbors. After exhaustive trials, k is set to 3. The testing set is given as input to the trained meta classifier (KNN) and final predictions are produced.

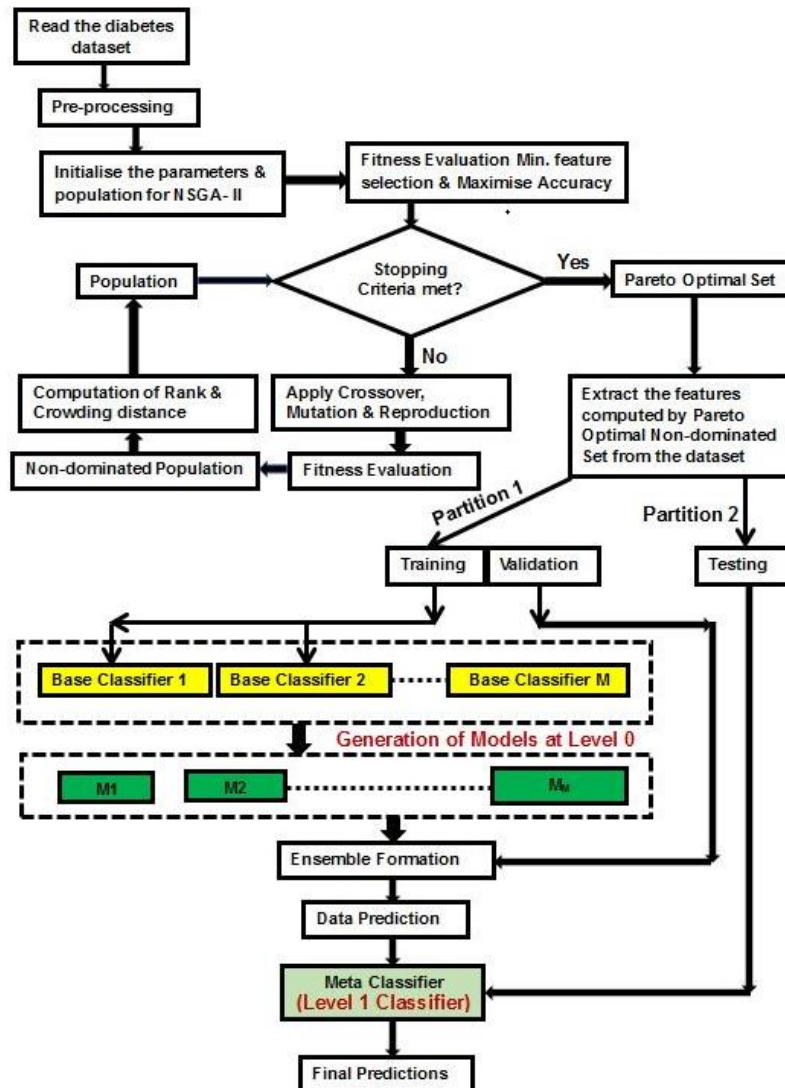


Figure 2. Proposed NSGA-II stacking model of T2DM prediction

5. EXPERIMENTAL DETAILS

In the study, 25 runs of the developed technique (NSGA-II stacking approach) were performed using MATLAB. Ballpark figure of 25 run was considered based on the results of the previous experimental work of the reference paper [9]. An extensive analysis involving comparison of the accuracy, specificity and sensitivity, error rate, of the proposed model with other approaches was done. Generalization performance of NSGA-II Stacking based predictive model of T2DM is included in this sub section. The graphs in Figures 3 and 4 represents trade off among the Pareto optimal solutions for error value-number of features selected from Pima dataset and collected dataset respectively.

The graphs drawn are for a one run where Y-axis denotes error value and X-axis represents number of features selected. Graphs in Figure 3 and 4 shows that each solution corresponds to number of features selected obtained by maintaining a balanced tradeoff between error value and number features selected. It has been observed that the error is minimum when 4 out of 8 features were selected in the Pima dataset while in the case of collected dataset the classification error was minimum when 20 out of 33 features were chosen.

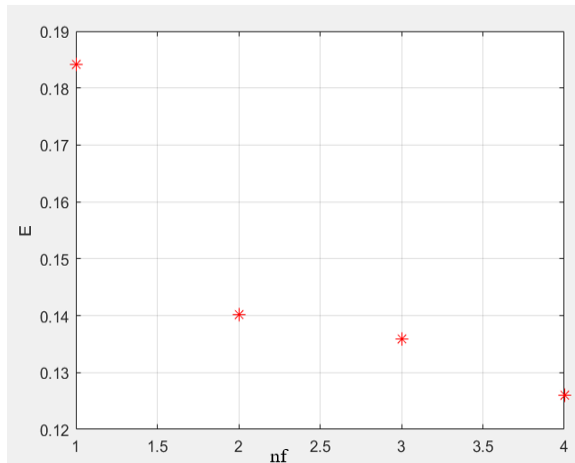


Figure 3. Error vs. number of attributes selected of the Pareto optimal solutions on testing set for one run on Pima dataset

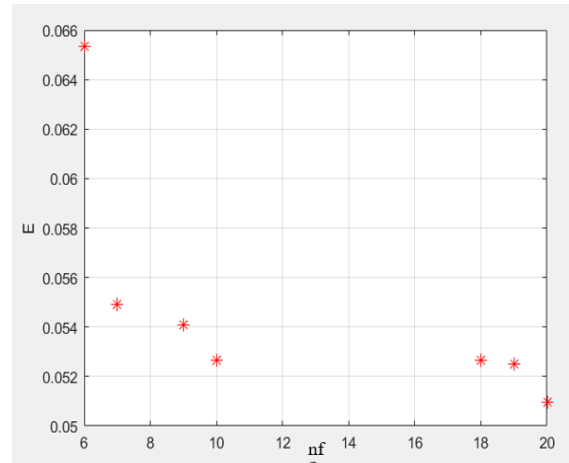


Figure 4. Error vs. number of attributes selected of the Pareto optimal solutions on testing set for one run on collected dataset

In the experiment performed, planned NSGA-II stacking approach has been compared with boosting, bagging, random forest and random subspace method. The performance of stacking approach has eclipsed the other conventional ensemble methods. It was noted that KNN has given a better performance over decision tree as a stacking combiner. Nine performance measures were computed for analyzing the classification performance of the NSGA-II stacking model. The parameters are accuracy of classification, precision, error, specificity, sensitivity, false positive rate, F1-score, Matthew’s correlation and kappa coefficients. These performance indicators are computed as listed in Table 1. TP, FP, TN and FN denote true positive, false positive, true negative and false negative respectively. The average model performance of NSGA-II stacking model is depicted in Table 2. Performance of NSGA-II stacking model on Pima and collected data is depicted in Figure 5.

Table 1. Performance indicators for assessment of model

Performance Indicators	Equations
Accuracy	$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + FP)}$
Error	$Error = \frac{(FP + FN)}{(TP + FP + FN + FP)}$
Sensitivity	$Sensitivity = \frac{TP}{(TP + FN)}$
Specificity	$Specificity = \frac{TN}{(TN + FP)}$
Precision	$Precision = \frac{TP}{(TP + FP)}$
False Positive Rate	$FPR = \frac{FP}{(FP + TN)}$
F1_score	$F1\ Score = \frac{2 * (precision * recall)}{(precision + recall)}$
Matthews Correlation Coefficient	$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (FP + TN) * (TP + FN) * (FN + TN)}}$
Kappa	$Kappa\ value = \frac{(observed\ accuracy - expected\ accuracy)}{1 - expected\ accuracy}$

Table 2. Performance of NSGA-II stacking model

Performance Parameters	Pima	Collected Dataset
Accuracy	81.90%	88.18%
Error	0.18	0.11
Precision	0.89	0.88
Specificity	0.80	0.94
Sensitivity	0.83	0.88
F1_score	0.86	0.87
Matthews Correlation Coefficient	0.61	0.82
False Positive Rate	0.19	0.05
Kappa Coefficient	0.61	0.74

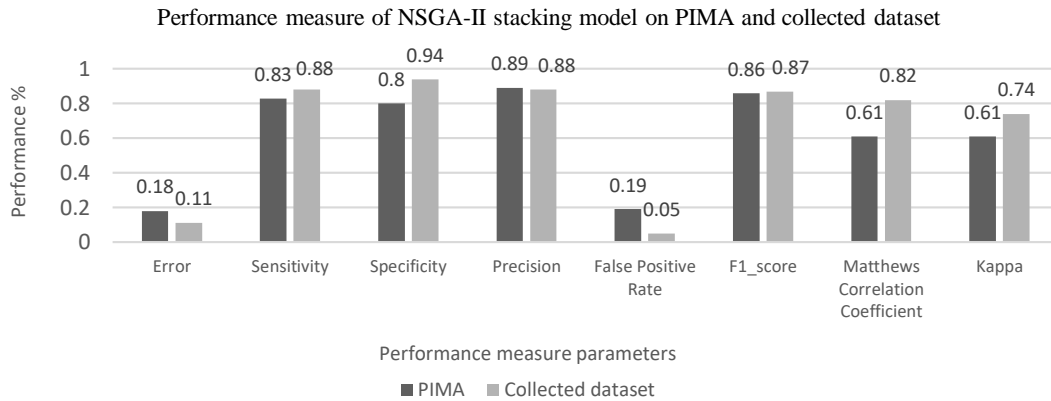


Figure 5. Performance of NSGA-II stacking model on Pima and collected dataset

Results of the suggested method is compared with the results obtained by other researchers since last five years. The benchmark dataset for the comparison with other models used is Pima by other researchers. Hence the proposed model was validated using the collected dataset and Pima dataset. Based on the prediction accuracy selected approaches from the current work are compared with the suggested model and shown in Table 3.

The comparison of the obtained predictive accuracy of proposed work with the existing methods over Pima has been shown in the Table 3. The result indicates that though the accuracy obtained was 81.9% over Pima dataset but the predictive model has outperformed over the collected dataset. The accuracy obtained was 88.18% over the collected dataset. After twenty-five runs of the experimental study, the chosen attributes varied but the mostly the attributes selected were weight, waist circumference, body mass index, HbA1c level, HDL and LDL cholesterol, VLDL, serum creatinine, Triglyceride level, fasting blood glucose (FPG) and post prandial glucose (PPG), family history, excess hunger, excess thirst, frequent urination, infection, poor wound healing.

Table 3. Results of existing T2DM predictive models in the literature (On Pima benchmark dataset)

Serial No	Methodology	Reference No.	Results on Pima (Accuracy %)
1	BPNN	[16]	81%
2	Decision tree	[17]	73.8%
	SVM		65.1%
	Naïve Bayes		76.3%
3	DNN	[1]	77.8%
	SVM		77.6%
4	Auto multilayer perceptron	[18]	88.7%
5	Stacking based Multiobjective evolutionary Ensemble	[10]	83.8%
6	PCA + ANN	[19]	75.7%
7	PCA + kmeans + LR	[20]	79.9%
8	PCA & minimum redundancy maximum relevance	[21]	77.2%
9	RMLP - Resampling version of MLP	[22]	79.3%
10	Gaussian fuzzy decision tree	[23]	75%
11	particle swarm optimization (PSO) with ANN	[24]	80%
12	Cultural Algo + ANN	[25]	79%
13	NSGA-II with stacking model (current work)	[16]	81.9%

Back propagation neural network (BPNN), deep neural network (DNN), principal component analysis (PCA), logistic regression (LR), and Resampling version of MLP (RMLP)




6. CONCLUSION

NSGA-II was utilized for feature selection and stacking approach was implemented for the prediction of T2DM. Stacking exploits the capabilities of a number of high performing models for classification and produces better results than any single model in ensemble. Extensive experimentation was carried out on the collected and benchmark dataset for verifying the efficacy of the developed models to find prediabetic stage. These models were compared with the existing ones and it was seen that the accuracy obtained was increased dramatically. The developed model (proposed) can detect this illness early (prediabetic stage), so that the physician and patient can work towards prevention and mitigation of complications caused by T2DM. The proposed work can be beneficial in myriad fields such as text mining, bioinformatics, image processing and for fault diagnosis and is of utmost significance for medical and data mining community. The proposed models' accuracy over Pima can be further improved by implementing outlier removal techniques.




REFERENCES

- [1] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, Feb. 2018, pp. 291–295, doi: 10.1109/WF-IoT.2018.8355130.
- [2] R. Rajni and A. Amandeep, "RB-bayes algorithm for the prediction of diabetic in Pima Indian dataset," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 4866–4872, Dec. 2019, doi: 10.11591/ijece.v9i6.pp4866-4872.
- [3] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12902-019-0436-6.
- [4] N. S. El Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," *Journal of Advanced Science*, vol. 124, pp. 1–10, 2018, doi: 10.14257/ijast.2018.124.01.
- [5] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
- [6] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020, doi: 10.1016/j.procs.2020.03.336.
- [7] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, Dec. 2021, doi: 10.1186/s42492-021-00097-7.
- [8] B. Farajollahi, M. Mehmannaavaz, H. Mehrjoo, F. Moghbeli, and M. J. Sayadi, "Diabetes diagnosis using machine learning," *Frontiers in Health Informatics*, vol. 10, no. 1, p. 65, Mar. 2021, doi: 10.30699/fhi.v10i1.267.
- [9] N. A. M. Aseri *et al.*, "Comparison of meta-heuristic algorithms for fuzzy modelling of COVID-19 illness' severity classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 50–64, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp50-64.
- [10] 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, Jan. 2020, doi: 10.1016/j.bbe.2019.10.001.
- [11] S. A. Diwan Alalwan, "Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 14, no. 1, pp. 88–95, Apr. 2019, doi: 10.11591/ijeecs.v14.i1.pp88-95.
- [12] S. Raghavendra and S. Kumar J, "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 353–359, Feb. 2020, doi: 10.11591/ijece.v10i1.pp353-359.
- [13] R. Sofiana and S. Sutikno, "Optimization of backpropagation for early detection of diabetes mellitus," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3232–3237, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3232-3237.
- [14] UCI Machine Learning, "Pima Indians Diabetes Database." Kaggle. Accessed: Aug. 21, 2021. [Online]. Available: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>
- [15] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, Apr. 2002, doi: 10.1109/4235.996017.
- [16] S. Joshi and M. Borse, "Detection and prediction of diabetes mellitus using back-propagation neural network," in *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, Sep. 2016, pp. 110–113, doi: 10.1109/ICMETE.2016.11.
- [17] 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.
- [18] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz, "An expert system for diabetes prediction using auto tuned multi-layer perceptron," in *2017 Intelligent Systems Conference (IntelliSys)*, 2017, pp. 722–728, doi: 10.1109/IntelliSys.2017.8324209.
- [19] T. Mahboob Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, 2019, doi: 10.1016/j.imu.2019.100204.
- [20] 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, 2019, doi: 10.1016/j.imu.2019.100179.
- [21] 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, Nov. 2018, doi: 10.3389/fgene.2018.00515.
- [22] N. MadhuSudana Rao, K. Kannan, X. Gao, and D. S. Roy, "Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution," *Computers & Electrical Engineering*, vol. 67, pp. 483–496, 2018, doi: 10.1016/j.compeleceng.2018.01.039.
- [23] K. V. S. R. P. Varma, A. A. Rao, T. Sita Maha Lakshmi, and P. V Nageswara Rao, "A computational intelligence approach for a better diagnosis of diabetic patients," *Computers & Electrical Engineering*, vol. 40, no. 5, pp. 1758–1765, Jul. 2014, doi: 10.1016/j.compeleceng.2013.07.003.
- [24] R. Patil and S. C. Tamane, "PSO-ANN-based computer-aided diagnosis and classification of diabetes," in *Smart Trends in Computing and Communications*, 2020, pp. 11–20.
- [25] R. Patil, S. Tamane, and K. Patil, "An experimental approach toward Type 2 diabetes diagnosis using cultural algorithm," in *ICT Systems and Sustainability*, 2021, pp. 405–415.




BIOGRAPHIES OF AUTHORS

Ratna Nitin Patil    holds a PhD in Computer Engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad in 2021. She also received her Bachelor of Engineering degree from Savitribai Phule Pune University, Maharashtra and completed Masters in Computer Engineering from Thapar University, Patiala, Punjab 1993 and 2001, respectively. She is currently an associate professor at Computer Science Department in Noida Institute of Engineering and Technology, Greater Noida, India. Her research includes meta-heuristics, machine learning and algorithms. She has published over 10 papers in international journals and conferences. She can be contacted at ratna.nitin.patil@gmail.com.






Shitalkumar Rawandale    holds a Doctor of Mechanical Engineering degree from RTM Nagpur University, Maharashtra, India in 2020. He also received his BE (Mechanical Engineering) and M.E (Mechanical Engineering) from North Maharashtra University, Maharashtra, India & from S P Pune University, Maharashtra, India in 1999 and 2004, respectively. He is currently working as Dean, Industry Institute Interaction at Pimpri Chinchwad College of Engineering, Pune, India since 2004. His research includes six sigma, mathematical models, prediction, and employability. He has published over 10 papers in international journals and conferences. He has also published 3 copyrights till now. He can be contacted at email: s.rawandale@gmail.com.






Nirmalkumar Rawandale    holds MBBS degree in 2002 and DNB in General Medicine 2008. He has vast experience of working in medical field and also in academics of medical education. He is well known for diagnosis and management of complicated diseases. He was the key person during Covid 19 pandemic management at GMC Dhule. He has worked as an Assistant Professor from 2009 to 2015, Associate professor of Medicine from 2015 and currently working as Head of Department of Medicine in Shri Bhausaheb Hire Government Medical College Dhule, Maharashtra, India. He is also a keen researcher in medical field with 7 publications in reputed international journals. He can be contacted at email: drrawandale@gmail.com.



Ujjwala Rawandale    is pursuing PhD from Savitribai Phule Pune University and obtained her Bachelor's degree in Electronics and Telecommunication Engineering from North Maharashtra University, Jalgaon, Maharashtra. She received her Master's degree from Prof. Ram Meghe College of Engineering, Amravati, India. Her area of interest is Signal processing. She is an Assistant Professor at Electronics and Telecommunication Engineering, MIT World Peace University, Pune. She is specialized in VLSI and Signal Processing, Speech Signal processing. She has published five papers at international level conference, and published four papers in the reputed journals. She can be contacted at email: ujjwala.rawandale@mitwpu.edu.in.



Shrishti Patil    is an MBBS student at Seth G.S. Medical College and King Edward Memorial Hospital. She is interested in pursuing interventional radiology. She was working on the front line during Covid 19 pandemic. She can be contacted at email: shrishtipatil1998@gmail.com.