International Journal of Pure and Applied Mathematics
Volume 118 No. 9 2018, 871-878
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm

Minyechil Alehegn
Symbiosis Institute of Technology
Pune, Maharashtra, India
Minyechil.tefera@sitpune.edu.in

Rahul Joshi& Dr. Preeti Mulay
Symbiosis Institute of Technology
Pune, Maharashtra, India
rahulj@sitpune.edu.in, preeti.mulay@sitpune.edu.in

Abstract **—Data mining techniques (DMTs) are very help full to predict the medical datasets at an early stageto safe human life. Large amount of medical datasets areopen in different data sources which used to in the real world application. Machine learning isa prediction on disease data. Currently, Diabetes Disease (DD) is the leading cause of death over all the world. To cluster and predict symptoms in medical data, various data mining techniques wereused by different researchers in different time. A total of 768 records, data set from PIDD (Pima Indian Diabetes Data Set) which is access from online source. In the proposed system mostknown predictive algorithms areapplied SVM, Naïve Net,DecisionStump, and Proposed Ensemble method (PEM).Anensemble hybrid modelbycombining the individual techniques/methods into one we made Proposed Ensemble method (PEM). Theproposed ensemble method (PEM) provides high accuracy of 90.36%**

Keywords—collaborative ; Diabetes; classification; Machine learning; Data mining;SVM,; Naïve Net;Decision Stump; PEM

I. INTRODUCTION

Currently in a global world, there are so many chronic diseases are distributed throughout the world, both in the developing and developed country such serious disease are distributed. From those serious diseases, Diabetes mellitus is one of the chronic diseases in the world which cut human life at early age. Diabetes Mellitus (DM) gets its name by health professionals' .At this time diabetes disease increases rapidly within the distance of light like Indian countries and some Saharan countries. It is not difficult to guess how much diabetes is very serious and chronic.There are different countries, organization, and different health sectors worry about this chronic disease control and prevent before the person died that means the early presentation of diabetes in order to save human life. Eating is also one factor for diabetes diseases and also, exercise used for healthy even a person live with diabetes the patient can recover from the disease by doing exercise Diabetes diseases have the power or ability to damage different parts of the human being body, from those human body parts which are affected by diabetes are listed as follow:-human heart, human eye, human kidney, and human nerves [39]. As it indicates it is easy to guess how much it is chronic and dangerous diseases that shorts human life. . Tao et al. [2] Algorithms which are used in machine learning have various power in both classification and predicting. Saba et al. [12] there is no single technique gives better performance and accuracy for all diseases, whereas one classifier provides or shows highest performance in a given dataset, another method or approach outdoes the others for other diseases. The new study or the proposed study concentrate on a novel combination or hybridization of different classifiers for diabetes Mellitus (DD) classification and prediction, thus overcoming the problem of individual or single classifiers. The new proposed study follows the different machine learning techniques (MLTs) to predict diabetes Mellitus (DM) at an early stage to save human life. Such algorithms are

SVM, Naïve Net, Decision Stump, and PEM to predict and increase the prediction accuracy and performance.

II. RELATED WORK

Song et al. [5]various algorithim was explained using different parameters such as Glucose, Blood Pressure (BP), Skin Thickness (ST), insulin, Body max index (BMI), Diabetes Pedigree function(DPF), and age. All parameters were not included.Only Small sample data used. ANN, EM, GMM, Logistic regression, and SVM were applied on diabetes dataset .ANN (artificial neural network) was provided better accuracy and performance than other algorithm.Xue-HuiMeng et al. [38] use different data mining techniques to predict the diabetic diseases using real world data sets by collecting information by distributing questioner. SPSS and weka tools were used for data analysis and prediction respectively. WeifengXu et al. [3] Different machine learning algorithm was applied in the prediction of diabetes diseases. From those algorithm RF was provided better accuracy than other data mining techniques. Loannis et al. [1] 10 fold cross validation was used as evaluation method in three different algorithms s Logistic regression, Naïve Bayes, and Svm. From those three different algorithm svm provided higher accuracy and performance than other method. Tao et al. [2] KNN, Naïve Bayes, Random Forest, decision tree, swim, and logistic regression was applied for the prediction purpose of diabetes mellitus (DM) at early stage. Concentrated on filtering. Yunsheng et al. [4] KNN and DISKR was used and storage space was reduced, an instance which has less factor was eliminated. Removing of outlier increase both performance and accuracy. Swarupa et al. [7].Naive Bayes (NBs) was providing good accuracy with the accuracy value of 77.01%. Sajida et al. [9] Adaboost provided better performance and accuracy. Pradeep &Dr. Naveen [8] J48 is one of the most popular and noted as better accuracy as well as good performance in this study.Pradeep et al. [11] J48 machine learning algorithm provided better performance and accuracy. Santhanam andPadmavathi [10] K-means, Genetic Algorithm, and SVM were applied and increase the accuracyvalue.Xue-Hui Men et al. [13] J48 was provided high performance. Croatia et al. [14] k-nearest neighbour (KNN) was applied and provided accuracy value of 70% accuracy. Ramiro et al. [6] wrongtreatment can reduced by applying fuzzy rule mechanismand also it helps for doctors as a recommender system in order to treat the patient without making mistake. Saba et al. [12] different data mining algorithm was applied from those algorithm Meta classifier provided higher accuracy than single classifier.

III. METHODOLOGY

In diabetic disease there were different research were done.sumarry of common or major findings are given as follow.

TABLE I. Summary of major findings or discoveries of diabetes prediction methodologies

Sn	**Author s**	**Methodologies**	**Finding**
1	Tao et al.[2]	KNN,Naïve Bayes, Decision Tree, Random Forest, SVM and Logistic Regression	Concentrated on the accuracy of recall and got better result. Filtering criteria can be improved
2	Loannis et al.[1]	Naïve Bayes, Logistic regression ,and Svm	From the three algorithm Svm provided high accuracy of 84%
3	WeifengXu et al.[3]	ID3 ,NaïvBayes,Randomforest,Adaboost	Random forest classifier method better relative to other .in contrast ID3 provided the least accuracy than others.
4	Yunsheng et al. [4]	DISKR and KNN	An attribute which have less factor should be eliminated. Accuracy increase can be increase by removing outliers. Spacecomplexity decreased.
5	Messan et al.[5]	GMM, ELM , ANN LR, and SVM	Less amount of sample data used. Comparison of algorithm were done from those method artificial neural network provide better accuracy than other classifier.
6	Ramiro et al.[6]	Fuzzy rule	Wrong treatment was reduced using fuzzy rule and recommendation system was developed for doctor.
7	Swarupa et al.[7]	,KNN,J48, ANN,zeroR, NB,cv parameter selection, Filtered classifier and simple cart	Various dataset appliedcontaining diabetes dataset. Cross validation not applied. NBshown high accuracy by providing accuracy of 77.01%.
8	Pradeep &Dr.Naveen [8]	Decision tree(J48)	J48 is noted as goodaccuracy provider algorithm. Feature selection has high role in the prediction area.
9	Sajida et al.[9]	Adaboost, j48,and Bagging,	Adaboost was shown improved accuracy than other method.

10	Santhanam and Padmavathi[10]	K-means with Genetic Algorithm ,and SVM	The integrated clustering and classification of algorithm done and provided better performance.
11	Pradeep et al.[11]	KNN, J48,SVM, and Random Forest	J48 providedefficient accuracy by providing 73.82% accuracy than others before pre-processing. Opposite side KNN and RF in provided good accuracy after pre-processing.
12	Saba et al.[12]	CART ,C4.5 ,Bagging, and ID3	The given algorithm applied on two diabetic datasets.
13	Xue-Hui Men et al.[13]	KNN, Logistic Regression, and J48	78.27% accuracy was measured using this method.
14	Krati et al.[14]	KNN	70% and 57% accuracy measured in data tes1 and data test2 respectively.
15	Saravananatha n and velmurugan[15]	SVM,CART, KNN,andJ48	j48, cart, svm and knn was applied and shown with the accuracy value of 67.15%, 62.28, 65.04 and 53.39 respectively.
16	Yang et al.[16]	NB, Bayes network.	72.3% accuracy measured by Bayes network
17	Asma [17]	Decision tree	Was shown 78.1768% accuracy.
18	Anjli and Varun[18]	SVM	72% accuracy measured
19	Thirumal et al.[19]	C4.5,SVM,KNN, and Naïve Bayes	C4.5 was shown improved accuracy than other with accuracy value of 78.2552%
20	Ayush and Divya[20]	CART	Was provided accuracy of 75%
21	Veena and Anjali[21]	SVM, Decision Stump,NB, and decision tree	80.72% accuracy was measured as better by Decision stump
22	Anuja and Chitra[22]	SVM	Betterperformance shownwith the accuracy value of 78% by this technique.
23	Prajwala[23]	DT and RF	Random forest was show better performance than decision tree
24	Bum et al.[24]	NB ,Logistic regression, and ,Anthropometry	Focused on prediction of Fasting Glucose Level. 74.1% performance and accuracy measured by anthropometry.
25	Aruna and Nazneen[25]	fuzzy rule, GA, and KNN,	Some rule was generated.
26	Sakorn[26]	Expert system and fuzzy rule	Focused on Expert system and was developed for treatment purpose.
27	Seokho et al.[28]	SVM ,E^2_SVM	80 % accuracy measured as better by using E^2_SVM
28	Emrana et al.[11]	KNN,C4.5	C4.5 provided more accuracy of 90.43 % and KNN provided accuracy of 76.96%
29	Kamadi et al.[30]	DT, Gini index, Gaussian fuzzy function	DT model shown good and efficient accuracy than other methods.
30	MunazaRamzan[29]	J48,Naïve Bayes ,and RF	RF provided better accuracy than J48 and Naïve Bayes in 10 cross validation Evaluation method.
31	Patil et al.[32]	HPM	92.38% accuracy recorded using HPM.
32	Abdullah et al.[31]	Support vector machine	Effective treatment of prediction was done using this technique.
33	Mounika et al.[32]	ZeroR,NB, and oneR	Effective treatment was applied on young and old patient. NB was better performance than others method
34	Nongyao and Rungruttikarn[34]	LR, Boosting, Naïve Bayes, ANN, Bagging, and Decision tree.	85.558% accuracy was measured using this Random Forest technique.it is recorded as better accuracy
35	Amit and Pragati [36]	RF,MLP,C4.5,and Bayes Net	The combination of MLP+BayesNet were shown better accuracy of 81.89% and better than other classification algorithm
36	Saba et al.[35]	NB,HMV,RF,Adaboost, KNN, LR,and SVM	Focused on various diseases including diabetes studied .78.085% accuracy measured by HMV algorithm it is recorded as better accuracy
37	Rian and Irwansyah[37]	Fuzzy Rule	Different rule was generated which helpsfor earlydetection of diabetes.

IV. CLASSIFICATION AND PREDICTION METHODS

Based on the extent literature, we established on employing four most known prediction algorithm such asSupport vector machine (SVM), NaïveNet (NN, and DecisionStump (DS) classification algorithm and combined the prediction of them in to one to increase the prediction accuracy of the algorithm using base learner.

A. Support Vector Machine (SVM)

Support vector machine algorithm is one of the most popular and widely used machine learning techniques.

Step1:-First we must identify the right hyper plane

Step2:-After the first step the second step ismaximizing the distances between neighbour data point

Step3:-Add a feature

z=x^2+y^2.it indicates that svm solves such problem.

Step4:-Apply Svm classifier to classify the class .the class is binary

B. Naïve Nets (NNs)

The time complexity of this technique is short .computes based on possibility by using the probability formula. It used to maximize the probability of (C|F)

Maximization =PR (class | feature)

Step1: The data should be convert into frequency table
Step2: Find likelihood
Step3:-In third stepuse naïve Bayes equation. Here the prediction is done.

(C|F) means PR (class | feature)

C. Decision Stump (DS)

It is one of the most popular machine learning classification algorithm that used in single level impute value .most of the time it is appropriate for an ensemble method specially in boosting that is one of the reason .

D. Collaborative (Ensemble) model

In prediction purpose individual prediction algorithms are not provided better and efficient performance the collaborative approach solves the limitation of distinct classifiers to cop up the accuracy better by combining in to one. [12, 32]

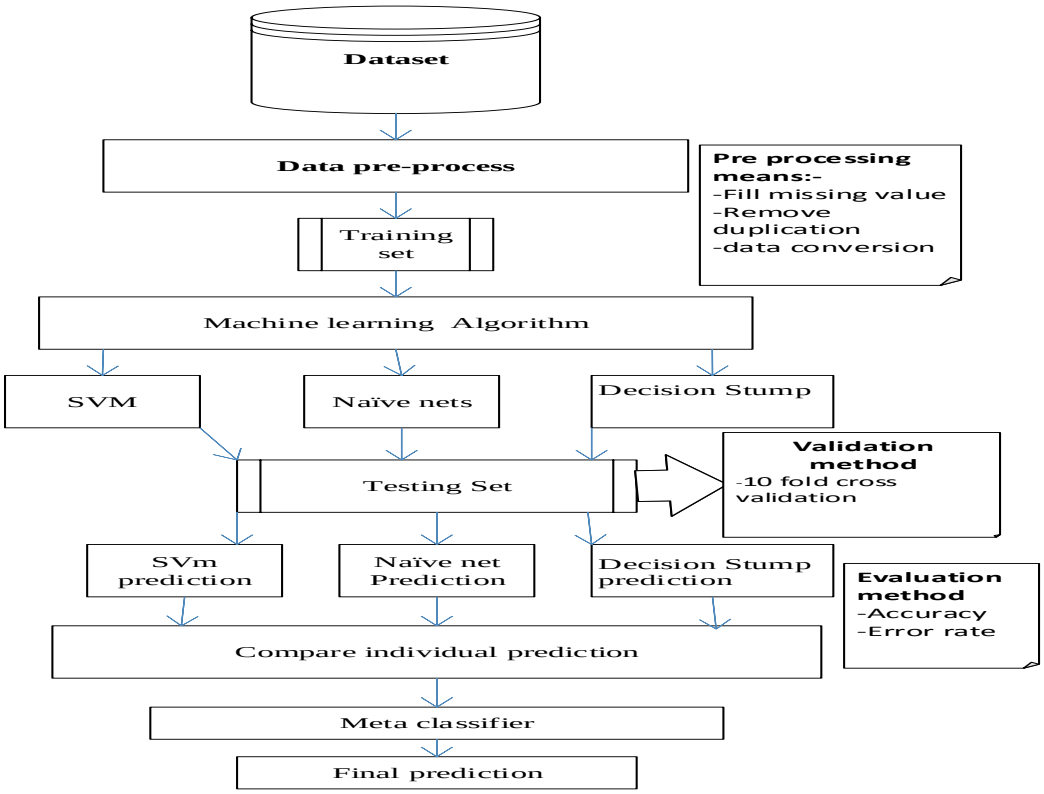

Fig 1: Proposed Work Flow

$$\textbf{Accuracy} = 100 * \left(\frac{\text{correctly claccified}}{\text{correctly classified} + \text{incorrectly classified}} \right)$$

TABLE II.The predictive accuracy (in percentage%) of the algorithm

Classification Algorithm	Accuracy (in %)	Incorrectly classified (in %)
SVM	88.8	11.2
Bayes Net	88.54	11.46
DecisionStumb	83.72	16.28
AdaBoostM1	85.68	14.32
Proposed method(PM)	90.36	9.64

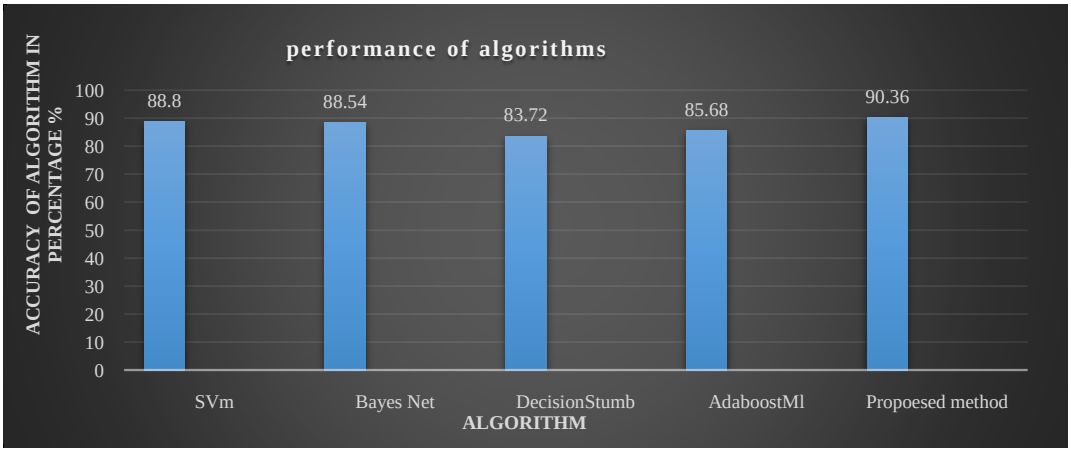

Fig 2. Accuracy of Algorithm

V. CONCLUSION

There are Various data mining method and its application were studied or reviewed .application of machine learning algorithm were applied in different medical data sets including machine Diabetes dataset.Machine learning methods have different power in different data set. We obtained 768record diabetes data set from UCI.the comparison of individual algorithm and the proposed method is done on this study. We applying 10 cross validation us for evaluation of the performance of these machine learning classification methods purpose. In this study the proposed method provide high accuracy with accuracy value of 90.36% and decisionStump provided less accuracy than other by providing 83.72% accuracy. Therefore, using ensemble method used to provide better prediction performance or accuracy than single one.

VI. FUTURE WORK

In this study we concentrated only Diabetes disease for future it can be extended to apply this method in another diseases Small amount sample data used on this study.it can be apply in large amount of data for future extension .on this study also only a single data set used therefore for future multiple data set can be used for prediction .in this study only limited base classifier used .for future it is possible to use another base classifier like ANN, Nave Bayes, KNN ,Random tree ,and other .

REFERENCES

[1] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." *Computational and structural biotechnology journal* (2017).

[2] Zheng, Tao, Wei Xie, LilingXu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." *International journal of medical informatics* 97 (2017): 120-127.

[3] Xu, Weifeng, Jianxin Zhang, Qiang Zhang, and Xiaopeng Wei. "Risk prediction of type II diabetes based on random forest model." In *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on*, pp. 382-386. IEEE, 2017.

[4] Song, Yunsheng, Jiye Liang, Jing Lu, and Xingwang Zhao. "An efficient instance selection algorithm for k nearest neighbor regression." *Neurocomputing* 251 (2017): 26-34.

[5] Komi, Messan, Jun Li, YongxinZhai, and Xianguo Zhang. "Application of data mining methods in diabetes prediction." In *Image, Vision and Computing (ICIVC), 2017 2nd International Conference on*, pp. 1006-1010. IEEE, 2017.

[6] Meza-Palacios, Ramiro, Alberto A. Aguilar-Lasserre, Enrique L. Ureña-Bogarín, Carlos F. Vázquez-Rodríguez, Rubén Posada-Gómez, and Armín Trujillo-Mata. "Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus." *Expert Systems with Applications* 72 (2017): 335-343.

[7] Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, pp. 1584-1589. IEEE, 2016.

[8] Pradeep, K. R., and N. C. Naveen. "Predictive analysis of diabetes using J48 algorithm of classification techniques." In *Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on*, pp. 347-352. IEEE, 2016.

[9] Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." *Procedia Computer Science* 82 (2016): 115-121.

[10] Santhanam, T., and M. S. Padmavathi. "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis." *Procedia Computer Science* 47 (2015): 76-83.

[11] Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." *Procedia Computer Science* 47 (2015): 45-51.

[12] Bashir, Saba, Usman Qamar, Farhan Hassan Khan, and M. YounusJaved. "An Efficient Rule-Based Classification of Diabetes Using ID3, C4. 5, & CART Ensembles." In *Frontiers of Information Technology (FIT), 2014 12th International Conference on*, pp. 226-231. IEEE, 2014.

[13] Meng, Xue-Hui, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors." *The Kaohsiung journal of medical sciences* 29, no. 2 (2013): 93-99.

[14] KratiSaxena, Dr, Zubair Khan, and Shefali Singh. "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm." *International Journal of Computer Science Trends And Technology (IJCST)* (2014).

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

[16] Guo, Yang, GuohuaBai, and Yan Hu. "Using bayes network for prediction of type-2 diabetes." In *Internet Technology And Secured Transactions, 2012 International Conference for*, pp. 471-472. IEEE, 2012.

[17] Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." In *Innovations in Information Technology (IIT), 2011 International Conference on*, pp. 303-307. IEEE, 2011.

[18] Negi, Anjli, and VarunJaiswal. "A first attempt to develop a diabetes prediction method based on different global datasets." In *Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on*, pp. 237-241. IEEE, 2016.

[19] Thirumal, P. C., and N. Nagarajan. "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study." *ARPN Journal of Engineering and Applied Science* 10, no. 1 (2015): 8-13.

[20] Anand, Ayush, and Divya Shakti. "Prediction of diabetes based on personal lifestyle indicators." In *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on*, pp. 673-676. IEEE, 2015.

[21] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." In *Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in*, pp. 122-127. IEEE, 2015.

[22] Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." *International Journal of Engineering Research and Applications* 3, no. 2 (2013): 1797-1801.

[23] Prajwala, T. R. "A comparative study on decision tree and random forest using R tool." *International journal of advanced research in computer and communication engineering* 4 (2015): 196-1.

[24] Lee, Bum Ju, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim. "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes." *IEEE journal of biomedical and health informatics*18, no. 2 (2014): 555-561.

[25] Pavate, Aruna, and Nazneen Ansari. "Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques." In *Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on*, pp. 371-375. IEEE, 2015.

[26] Mekruksavanich, Sakorn. "Medical expert system based ontology for diabetes disease diagnosis." In *Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on*, pp. 383-389. IEEE, 2016.

[27] Hashi, EmranaKabir, MdShahidUzZaman, and MdRokibulHasan. "An expert clinical decision support system to predict disease using classification techniques." In *Electrical, Computer and Communication Engineering (ECCE), International Conference on*, pp. 396-400. IEEE, 2017.

[28] Kang, Seokho, Pilsung Kang, TaehoonKo, Sungzoon Cho, Su-jin Rhee, and Kyung-Sang Yu. "An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction." *Expert Systems with Applications* 42, no. 9 (2015): 4265-4273.

[29] Ramzan, Munaza. "Comparing and evaluating the performance of WEKA classifiers on critical diseases." In *Information Processing (IICIP), 2016 1st India International Conference on*, pp. 1-4. IEEE, 2016.

[30] Varma, Kamadi VSRP, AllamAppa Rao, T. SitaMaha Lakshmi, and PV Nageswara Rao. "A computational intelligence approach for a

better diagnosis of diabetic patients." *Computers & Electrical Engineering* 40, no. 5 (2014): 1758-1765.

[31] Aljumah, Abdullah A., Mohammed GulamAhamad, and Mohammad Khubeb Siddiqui. "Application of data mining: Diabetes health care in young and old patients." *Journal of King Saud University-Computer and Information Sciences* 25, no. 2 (2013): 127-136.

[32] Patil, Bankat M., Ramesh Chandra Joshi, and DurgaToshniwal. "Hybrid prediction model for type-2 diabetic patients." *Expert systems with applications* 37, no. 12 (2010): 8102-8108.

[33] Mounika, M., S. D. Suganya, B. Vijayashanthi, and S. Krishna Anand. "Predictive analysis of diabetic treatment using classification algorithm." *IJCSIT* 6 (2015): 2502-2505.

[34] Nai-arun, Nongyao, and RungruttikarnMoungmai. "Comparison of classifiers for the risk of diabetes prediction." *Procedia Computer Science* 69 (2015): 132-142.

[35] Bashir, Saba, Usman Qamar, Farhan Hassan Khan, and Lubna Naseem. "HMV: a medical decision support framework using multi-layer classifiers for disease prediction." *Journal of Computational Science* 13 (2016): 10-25.

[36] kumarDewangan, A., & Agrawal, P. (2015). Classification of Diabetes Mellitus Using Machine Learning Techniques. *International Journal of Engineering and Applied Sciences*, *2*(5), 145-148.

[37] Lukmanto, Rian Budi, and E. Irwansyah. "The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model." *Procedia Computer Science* 59 (2015): 312-319.

[38] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, *29*(2), 93-99.

[39] *http://www.who.int/mediacentre/factsheets/fs312/en/*

[40] *https://www.kaggle.com/uciml/pima-indians-diabetes-database*