Diabetes Analysis And Prediction Using Random Forest, KNN, Naïve Bayes, And J48: An Ensemble Approach

Minyechil Alehegn, Rahul Raghvendra Joshi, Preeti Mulay

Abstract:Now-a-days there is increase in people suffering from DM (Diabetes mellitus) and this number is growing continuously. So, it is a considerable chronic disease. MLTs (Machine Learning Techniques) can act as a savior for early diagnosis and prediction of DM. ML is another side of Artificial Intelligence so that be used for prediction, recommendation and recovery from disease in early stages. The system proposed in this paper makes use of two datasets viz. PIDD (Pima Indian Diabetes Dataset) and 130_US hospital diabetes data sets. Techniques used for datasets analysis are Random Forest, KNN, Naïve Bayes, and J48. Ensemble approach facilitates in achieving better results. The accuracy of proposed ensemble approach is 93.62% for PIDD and 88.56% for 130_US hospital dataset.

Index Terms:J48, Diabetes, Ensemble KNN, Naïve Bayes, Random forest,

1 INTRODUCTION

Diabetes usually known as DM. It is a kind of metabolic diseases in which patients suffer from blood glucose problems due to abnormal production and release of insulin. As per WHO report on 14th November 2016, i.e. on World Diabetes Day, 422 million adults are living with diabetes, and 1.6 million people who lost their life due to DM [1]. In 2016, 1.6 million deaths were directly caused by diabetes [1]. So, it is one of the seriously need to be considered kind of chronic disease around the world. DM can cause damage to different body parts viz., nerves, eyes, heart, to name a few. Every year millions of people got affected by this life threatening disease in both civilized and non-civilized parts of the world. CDCP (Center for Disease Control and Prevention) projected that during 2001 to 2009 there is 23% increase in Type II diabetes in US [2]. Many countries, Organization, and different health sectors are also worried about this chronic disease for achieving control and prevention in order to mitigate it in early stages, so that person life can be saved. Different variants of DM are there viz. Type I, Type II, Juvenile and Gestational. Type I is insulin dependent, Type II is insulin independent, Gestational can happen during pregnancy and Juvenile diabetes after birth of a baby. According to Canadian Diabetes Association (CDA) in coming 10 years that is during 2010 to 2020, there will be a predictable growth from 2.5 to 3.7 million for people suffering from chronic diseases [3]. So, by looking at these statistics diabetes and other chronic diseases analysis plays a vital role in saving patients life. Moloud et al. [4] discussed ML algorithms used for analysis and prediction purpose will have different processing powers. Meng, Xue-Hui, et al. [5] showed that ML methods/tactics are helpful in getting better insights, patterns from input health data. Bashir, Saba, et al. [6] confirmed that single machine techniques are less effective as compared distributed implementation as well doesn't work well for single dataset. So, in this paper proposed approach makes use of two datasets i.e. PIDD and 130_US hospital diabetes datasets. Also, ML techniques used here are Random Forest, KNN, Naïve Bayes, and J48 etc. This paper spans over five sections viz., related work, methodology, proposed prediction and classification along with results, conclusions. References used in this paper are consolidated at the last.

2 RELATED WORK

Song et al. [7] explained and described using various factors such as Age, Glucose, BP, BMI, Skin Thickness etc. Diabetes Pedigree function, insulin, and pregnancy parameters not included [7]. Small sample data used in this study for the prediction of DM. Algorithms used were EM, LR, GMM, SVM and ANN. ANN showed high accuracy and performance [7]. Loannis et al. [8] used Naïve Bayes, SVM, and LR. 10 fold cross validation evaluation method was applied. These individual methods compared based on their accuracy value. Among these three algorithms SVM provides visible result than that of other prediction. Data origin, its kind, and dimensionality are some of the factors related to accuracy. Nilashi et al. [9] for noise removal and pre-processing used two clustering techniques viz., PCA and EM. The medical datasets used were Diabetes, Heart, etc. Removal of noise from instances helped to get better accuracy. Yunsheng et al. [1] removed parameters having less influence which in a way can increase or improve performance and gives better outcome .Francesco et al.[10] to escalate the accuracy of algorithm used feature selection mechanism which plays a great role in improvisation of both accuracy as well as performance. HT, DT, BN, MP, Jri, and RF methods were used for analysis and prediction. Best first (BF) and greedy used for feature selection. Hoeffding Tree was shown good performance and better accuracy. Pradeepet al. [11] not used anycross validation. J48, NBs, Cv Parameter selection (CPS), Simple Cart (SC), ANN, ZeroR, filtered classifier, and KNN techniques were applied. Naïve Bayes provide better accuracy in case of DM than other techniques. ANN and KNN gave accurate result for other datasets than that of other prediction algorithms. Sajida et al. [12] used three technique viz. Bagging, J48 and

- *Minyechil Alehegn is faculty at IT department of Mizan Tepi University, Ethiopia.*
- *Rahul Raghvendra Joshi and Dr. Preeti Mulay are faculty at CS/IT department of Symbiosis Internal (Deemed University), Pune, Maharashtra, India.*

Adaboost on CPCSSN dataset to predict the DM at early stage to prevent from early death. Adaboost technique shown effective and improved result compared to other data mining methods. Kamadi et al. [13] data reduction is the problem in cased of classification and it also plays a great role in the prediction. To get good accuracy data need to be reduced. PCA does pre-processing. So, for getting better results data need to be reduced. Pradeep & Dr. Naveen [14] showed accuracy of the prediction algorithm is different before and after pre-processing. Decision Tree Technique showed good accuracy before pre-processing the DM dataset. Two algorithms Random Forest and Support Vector Machine does better after pre-processing of DM dataset. Santhanam and Padmavathi [15] dimensionality reduction plays an important role to improve accuracy. K-Means and Genetic Algorithm used to reduce dimensions. 10K folds cross validation mechanism used for evaluating the method. Un-reduced data proved to be less accurate than that of reduced data. Xue-Hui Meng et al. [16] data can be collected by different means like by using distributed questioner also; LR, J48, and ANN. Decision Tree classification technique gave better accuracy than that of ANN and LR. Abdullah et al. [17] OD (Oracle Database) and ODM (Oracle Data Miner) used for storage analysis. Target variables can be identified based on their percentage of necessity. Malgorzata et al. [18] validity of population based forecast was studied using Closed Loop (CL). Juliahippisley et al. [19] used QD score and 5K cross validation methods to get better result. Philippa et al. [20] in his study the most important factors related to Type II diabetes were identified by using Genetics Score (GS) data mining or machine learning algorithm. Robert et al. [90] used CACS and got better results. Pérez et al. [21] ANN, ML or Data Mining prediction algorithms or techniques were applied. ANN replaces human brain with new technology. It is somewhat complex and difficult to for fresher to work out. Results of this predictor are better in almost all cases. Kazuaki et al. [22] relationship or association of genes was identified using LR method. Muhammad et al. [23] identified risk factors using association rules. E. S. Kilpatrick et al. [24] mean blood glucose risk factor showed best predictor than that of HbA1c in Type I diabetes using Cox Regression (CR).

3 Findings from Existing Literature related to DM Prediction

ML or Data Mining as well as AI with health care industry are doing well in terms prediction. Findings from existing literature related to DM prediction are as follows:

Sr. No.	Authors	Methodologies	Findings
1	Weifeng Xu et al.[25]	Adaboost,ID3.NB, and RF	Better accuracy by RF and less accuracy by ID3
2	Messan et al.[26]	LR,GMMANN,SVM, ELM	ANN was best with 89% accuracy.
3	Loannis et al.[8]	LR, NB, SVM	84% accuracy by SVM with 10K cross validation.
4	Mehrbakhsh et al.[9]	PCA, EM, CART	Noise removal played an important role.
5	Tao et al.[27]	NBs, KNN, J48, RF, LR, and SVM	Improvisation in cleansing improved system performance.
6	Yunsheng et al.[7]	KNN, DISKR	Less important attributes and outlier were removed for obtaining better results.
7	Francesco et al.[10]	Hoeffding Tree, Decision Tree, ANN, Jrip, Bayenet, Greedy Stepwise, Best First, and RF	Main focus was on feature selection (FS). 10K fold cross validation evaluation method was applied. Hoeffding Tree (HT) showed high accuracy with integration of searching algorithm achieved accuracy of 77.5% than that of other method.
8	Swarupa et al.[28]	NB, ANN, KNN, J48, zeroR, CPS, SC, and FC	Cross validation mechanism not applied. NB showed relatively good performance of 77.01%.
9	Sajida et al.[12]	Decision Tree, Bagging, and Adaboost	Adaboost showed better performance than other considered method
10	Munaza Ramzan [29]	J48,NB,and RF	RF showed improved accuracy than other techniques.10 fold cross validation used was effectual in this study.
11	Kamadi et al. [13]	Modified fuzzy and PCA	Data reduction improved overall accuracy.
12	Pradeep & Dr.Naveen [29]	J48	Feature selection technique used to improve correctness. Decision tree noted as a better performing algorithm.
13	Ramiro et al.[30]	Fuzzy Rule	Recommender system decreased wrong treatment.
14	Pradeep et al.[11]	RF, KNN, Decision Tree, KNN, RF, and SVM	Before pre-processing DT showed accuracy of 73.82%. RF and KNN showed also good performance after pre-processing.
15	Santhanam and Padmavathi [15]	K-Means, GA, and SVM	Hybridization of classification and clustering algorithms improved accuracy.
16	Sankarana &Dr Pramananda [31]	Association Rule using FP growth and Apriori	Easy and simple decision making helped, acted as a defensive and suggestive medicine.
17	Xue-Hui Men et al.[5]	KNN, LR, and J48	J48 gave accuracy of 78.27%.
18	Abdullah et al.[17]	SVM	SVM predicts focused treatment.
19	Patil et al.[32]	HPM	It gave high accuracy of 92.38%.
20	Saba et al.[6]	Adaboost, RF, KNN, LR, NB, SVM, and HMV	Different diseases were studied and output of HMV was attractive.
21	Amit and Pragati [33]	MLPBayesnet, C4.5, and RF	MLP+Bayesnet showed better performance.
22	Saba et al.[34]	C4.5, Bagging, ID3, and CART	Bagging classifier showed high accuracy than that of other

			methods.
23	Mounika et al.[35]	NB, ZeroR, and oneR	Effective treatment to both old and young patients. NB showed good performance than that of others.
24	Nongyao and Rungruttikarn [36]	NB, LR, Bagging, ANN, Boosting, and J48.	The concept applied by using boosting or bagging. RF showed accuracy of 85.558%
25	Dr Saravana et al.[37]	Analysis algorithm in Hadoop	Focused on treatment in healthcare industry using big data analysis. Low cost for proper treatment.
26	Veena and Anjali [38]	Decision Stump(DS), SVM, J48, and NB	DS (Decision stump) algorithm provided better accuracy of 80.72%.
27	Kung et al.[39]	KNN, New EM, and opposite sign test	The hybridization of EM and KNN used in this study.
28	Saravananath an and velmurugan [40]	SVM, J48, CART, and KNN	J48 prediction showed better output with accuracy of 67.15%.
29	Seokho et al.[41]	SVM, E2_SVM	It was an ensemble approach. E2_SVM was exposed better prediction than normal Support vector machine with 80 %.
30	Rian and Irwansyah [42]	Fuzzy rule	Early detection is done by fuzzy rule.
31	Yang et al. [43]	NB, BN.	BN showed improved accuracy of 72.3%
32	Lin [44]	NB, SVM, ANN	The meta classifier used in this study. By applying meta classifier the accuracy of the algorithm can be improved.
33	Vrushali and Rakhi [45]	CLAT	Estimation of Prediction and severity to diabetes can be accurately analyzed.
34	Emrana et al.[46]	C4.5 and KNN	C4.5 showed results with accuracy of 90.43%.
35	Nahla et al[47]	SVM with rule extraction with SQRex-SVM	Hybridized method.
36	Kamadi et al.[48]	DT, and Gini index, Gaussian fuzzy function	DT model showed better performance.
37	Sakorn [49]	Expert system with fuzzy rule	Expert system for better treatment developed.
38	Ayush and Divya [15]	CART	Study showed 75% of accuracy.
39	Jae et al. [50]	Linear and Wrapper selection	Time complexity was minimized.
40	Bum et al.[51]	LR and Naïve Bayes, Anthropometry	Study was concentrated on forecast of glucose levels. The results showed 74.1 % accuracy.
41	Asma [52]	Decision Tree(DT)	DT showed results of 78.1768%.
42	Anjli and Varun [53]	SVM	RF and SVM achieved 72% of accuracy.
43	Aruna and Nazneen [54]	GA,KNN, and fuzzy rule	Association between GA and KNN were done. Fuzzy rule developed in this study.
44	Prajwala [55]	Random Forest and Decision Tree	RF prediction showed better accuracy than DT
45	Emirhan et al.[43]	Rough Set, ANFIS	ANFIS was important and showed improved performance than Rough Set.
46	Krati et al. [56]	KNN	Dependency on dataset size.
47	Anuja and Chitra [57]	SVM	SVM prediction method provided 78% of accuracy.
48	Thirumal et al.[58]	C4.5, KNN, NB, SVM	C4.5 provided better accuracy of 78.2552% as compared to others.
49	E. S. Kilpatrick et al.[59]	MBG, HBAc1	MBG was better prediction algorithm than that of HBAc1
50	N. Sattar et al .[60]	Association	Novel factors for DM were identified
51	C. Törn et al.[61]	GAD	GAD performs better and gives proper output.
52	M. S. Lewi et al.[62]	IGFBP-1	Glucose level is one of the prominent factor for DM.
53	H. Elding Larsson et al.[63]	SDS (Standard Deviation Score)	Birth date is also associated with diabetic risk.
54	K. Chien et al.[64]	Cox Regression Coefficients(CRC)	Prediction model developed 5K folds cross validation used.
55	R. W. Grant et al.[65]	Genetic test	Hypothesis formulated for Type II DM.
56	Y. Vergouw et al.[66]	MLR	Rules were generated for Type I diabetes.
57	M. Ekelu et al .[67]	HOMA-beta, HOMA	Gestational diabetes occurs only at the pregnancy time.
58	B. Isom et al.[68]	OGTT	Family history is one of the factors related to diabetes.
59	matthew l et al.[69]	Risk Adjustment Technique, R2	DM specific measures done in this study.
60	Antti-jussi pyykkonenetal .[68]	Association	Life style leads to poor health.
61	Dan ziegler et al.[70]	Cox Proportional Hazard Models (CPHM)	Predictors of mortality for diabetic & non diabetic population studied.
62	Kyoko kogawa satoet al.[71]	LR	The combined of A1C and fast plasma glucose studied.
63	Alison jet al.[72]	CPHM	Self rated health profile combined with EQ VAS was an important factor.
64	Alehegn M et al.[73]	SVM, NN, DS, and PEM	PEM showed better performance.
65	NAnnekleefstr a et al.[74]	Association	Better living style and Type II diabetes are related to each other.
66	Meredith et al[75]	Logistic Regression Models (LRM)	Waist-to-height ratio is the most predictive measure.
67	Mikael et al.[76]	Mean Square(X2)	Type I diabetes analysis was more effectual using Mean Square method.
68	caroline et	Regression Model	Uric acid was better

	al.[77]		predictor for DM.
69	sylvia et al.[78]	Association	LB has direct relationship with diabetes as it was confirmed. So, it is one of the factors related to rise of the diseases.
70	Wei et al.[79]	SVM	Sex and physical activity are considerable attributes for DM.
71	Mohammadreza et al.[80]	Operating Characteristic Curves	Lipid accumulation product was important predictor for DM.
72	Farzad et al.[81]	SD	HDLC foretelling exists in young women rather than male.
73	michael et al .[82]	Archimedes Model (AM)	AM predicted DM with better accuracy and high level of sensitivity.
74	Earl et al.[83]	Association	Metabolic syndrome is the strong cause for occurrence of diabetes.
75	Marion et al.[84]	Logistic Regression (LR)	Females living with Type I diabetes studied.
76	Jan cederholm et al.[85]	Cox Regression Analysis(CRA)	Type II DM studied only in this study.
77	Amber et al.[19]	Score, Ukpds, Framingham algorithm	Ukpds and score were useful than Framingham algorithm.
78	Matthias et al.[86]	Operating Characteristic Curve	Life style is one of the factors for DM occurrence.
79	Heli et al. [87]	Association	Association of HLA and autoantibody was studied
80	Makrina et al.[88]	Integrated Discrimination Improvement (IDI)	Lifestyle distinctions may can't only reduce diseases possibility for a women who were not pregnant
81	Mandy et al.[61]	LR	High risks related to DM studied.
82	Rachel et al[89]	Bergman Model	This model gave better outcome.
83	Daniel et al.[90]	ARX model and model-free ZOH	ARX was better than ZOH in this study.
84	Malgorzata et al.[18]	Closed Loop (CL)	Validation for population based forecast was studied
85	Juliahippisley et al. [91]	QD Score	Five cross validation was used to achieve result.
86	Philippa et al.[20]	Genetics Score (GS)	Most of the risk factors related to Type II DM were identified.
87	Robert et al.[85]	CACS	CACS predictor showed better results.
88	Pérez et al.[21]	ANN	ANN outperformed for DM analysis.
89	Kazuaki et al.[22]	Logistic Regression(LR)	The genes association was studied and identified in relation to DM.
90	David G. Clayton [92]	Association	Type I DM causes studied.
91	Muhammad et al.[93]	ANOVA	Possibility of Type II DM for future generation was analyzed.
92	Abu nasar et al.[94]	CLIPS	Expert system for plant diseases developed.
93	Orhan et al.[95]	ANN	ANN outperformed as compared to other DM predications.
94	Esin Dogantekin et al.[96]	LDA-ANFIS, ANFIS	LDA-ANFIS provided better accuracy of 84.61% which is much better than ANFIS.
95	Muhammad et al.[23]	Association rule	DM risk factors identified.
96	E. S. Kilpatrick et al.[24]	Cox regression	Mean blood glucose risk was best predictor than HbA1c in Type I diabetic
97	dan ziegler et al.[92]	Association	Risk relation to disease possibility was studied.
98	Törn et al.[97]	IA-2A ,Roc ,GADA	Roc was less significant as compared to others.
99	Alehegn M et al.[73]	SVM, NN, DS and PEM	PEM showed effectual results.

4 Proposed Prediction and Classification Method

The dataset from UCI repository i.e. PIDD and 130-US hospital dataset were considered [54]. PIDD involves 768 records and 8 characteristics with one target class and 130–US hospital dataset consists of 93743 instances and 48 features. Ensemble or hybrid model with base learner for forecasting applied. Based on literature, four most known data mining or machine learning prediction techniques were considered. They are described as below.

4.1 RANDOM FOREST (RF)

It is one of the prediction algorithms in the machine learning area. It is more adaptable to ensemble approach. It can easily tackle large datasets.

4.2 K –NEAREST NEIGHBOUR (KNN)

It is grouped under the category of lazy prediction technique. It is easy technique helps to group new work based on similarity measure [18]. The training data are sorted in this algorithm. Define k - number of nearby neighbours. Distance between training samples and instance. Estimate inaccessibility of the training sections arranged and the neighbouring neighbour based on the minimum - the remoteness is determined in the subsequent step. Training data for all categories defined. Majority of the class of nearest neighbours have the forecast value of the query record. Euclidean= $\sqrt{\sum_{i=1}^{k}(xi - yi)}$

4.3 NAÏVE BAYES (NB)

NB is prevalent and fits when the input data is large and need a short computational time. Calculation based on prospect is done by applying Bayes formula [19]. $p(h/D) = (p\ (D/h)\ p(h))/(p(D))$ Where P (h) is refers to prior probability of hypothesis, h in this case is true P (D) is refers to prior possibility of training data D P (h/D) is refers to possibility of h given D P (D/h) is refers to possibility of D given h

4.3 J48 (DECISION TREE)

It is also called decision tree prediction algorithm. It is the upgraded version of ID3 classification machine learning algorithm. By using this algorithm, it is possible to construct rules which are simple and easy to understand [47]. Check for the above base cases. For each Instance I, find the consistent information gain ratio by splitting on I. Let I_better be the feature with the maximum or better normalized information gain. Make a choice node that splits on I_better. Recur or repeat on the sub lists obtained by splitting on I_better, and add those bulges as leaf of node.

4.5 HYBRID MODEL

By combining 4.1 to 4.4 to one method – accuracy will be maximum [12, 47].

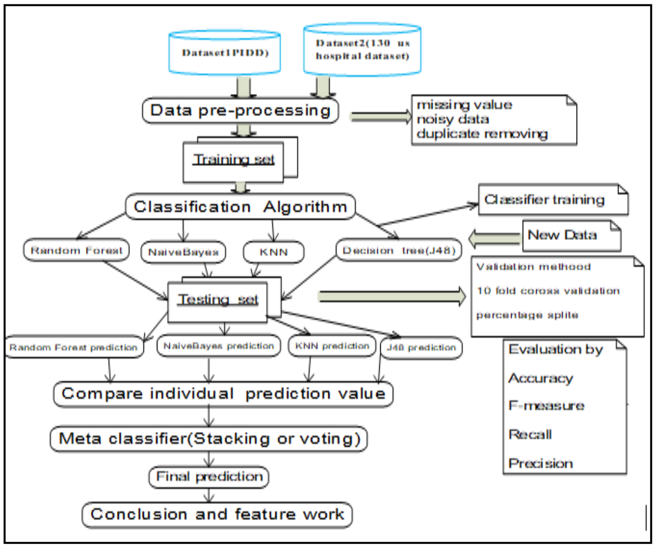

Fig 1: Proposed Work Flow

5 EQUATIONS

Table 2 Accuracy of PIDD against considered Algorithms

Classification Algorithm	Correctly Classified	Incorrectly classified
J48	87.89%	12.11%
KNN	82.94%	17.06%
Naïve Bayes	88.41%	11.59%
Random Forest	89.84%	10.16%
Proposed Ensemble using stacking	**93.62%**	**6.38%**
Baging-J48	89.97%	10.03%
Bagging-KNN	85.81%	14.19%
Bagging-Random forest	89.93%	10.07%

TABLE 3 ACCURACY OF 130_US AGAINST CONSIDERED ALGORITHMS

Classification Algorithm	Accuracy before fs	Incorrectly classified	Accuracy after fs	Incorrectly classified
J48	56.96%	43.04%	84.53%	15.47%
KNN	46.04%	53.96%	74.59%	25.41%
Naïve Bayes	56.18%	43.82%	82.26%	17.74%
Random Forest	48.68%	51.32%	82.80%	17.20%
Proposed Ensemble using stacking	**58.04%**	**41.96%**	**88.56**	**11.44%**
Baging-J48	57.06%	42.94%	86.96%	13.04%
Bagging-KNN	50.76%	49.24%	77.89%	22.11%
Bagging – Naïve Bayes	53.76%	46.24%	84.61%	15.39%
Bagging-Random forest	54.55%	45.5%	85.15%	14.85%

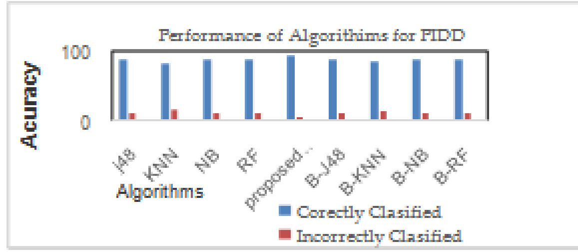

Fig 2: Accuracy of considered of PIDD for different algorithms

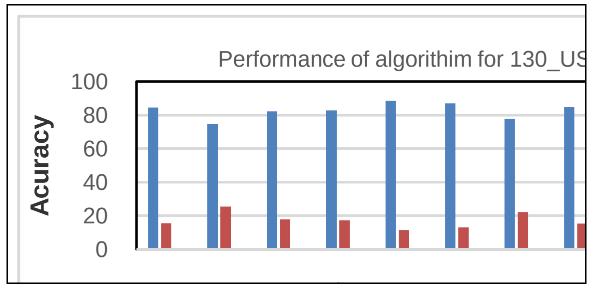

Fig. 3: Accuracy of us-130 hospitals dataset for different algorithms

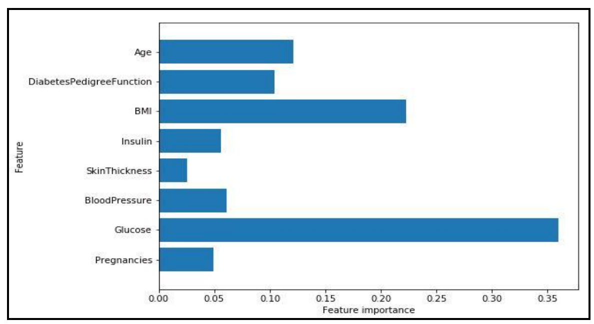

Fig 3: Important Features of PIDD

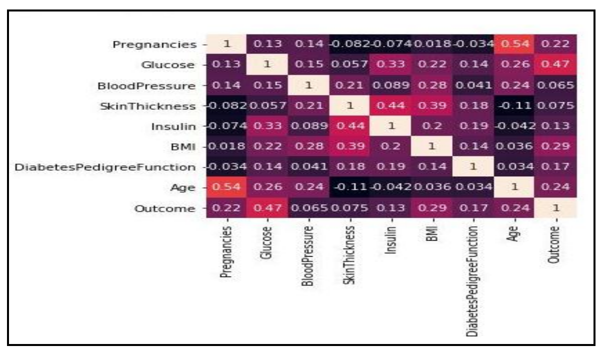

Fig. 4: PIDD Attributes correlation

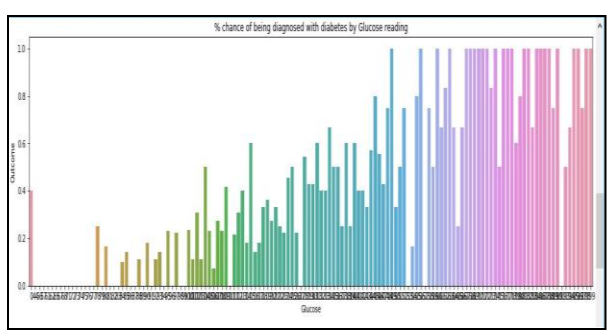

Fig. 6 Chance of being diagnosed for DM by Glucose

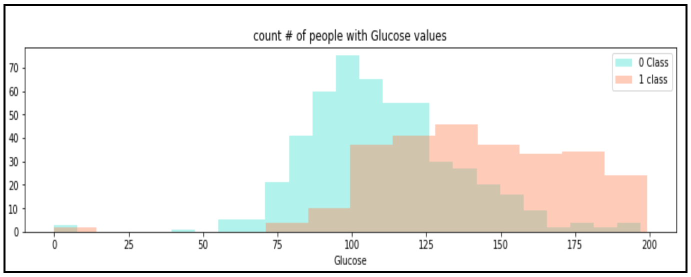

Fig. 7 Chance of being diagnosed for DM by age

6 Conclusions

Deep Learning or Machine Learning methods have different powers for diverse data sets. In the proposed system two datasets one is large (130_US) and other is small (PIDD) used for analysis. In this work 10K cross validation for evaluation both in single and multiple iterations applied by considering 90% of training and 10 % of testing data. Proposed system used well known and most commonly used machine learning algorithms. Algorithms used in this study are J48, KNN, NB, and Random Forest. The proposed method provides better accuracy of 93.62% in case of PIDD using stacking meta classifier. In case of large dataset 130-us hospital an ensemble method provides better accuracy than single prediction algorithm. Generally in both small and large datasets analysis ensemble method outperformed than a single method. It is also observed that when dataset becomes large the accuracy of the proposed algorithm is not good relatively. NB and J48 prediction algorithm are better for large datasets analysis. KNN technique is not good for large dataset analysis. In this study focus is on DM analysis, in future this hybrid approach or ensemble approach needs to be applied on other diseases for gauging its effectuality.

References

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes accessed on 20th July 2019.
[2] https://www.cdc.gov/media/releases/2017/p0718-diabetes-report.html accessed on 20th July 2019.
[3] https://www.diabetes.ca/ accessed on 20th July 2019.
[4] Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251.
[5] 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.
[6] Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: A medical decision support framework using multi-layer classifiers for disease prediction. Journal of Computational Science, 13, 10-25.
[7] Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing, 251, 26-34.
[8] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.
[9] Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 106, 212-223.
[10] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia computer science, 112, 2519-2528.
[11] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.
[12] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82, 115-121.
[13] Kamadi, V. V., Allam, A. R., & Thummala, S. M. (2016). A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach. Applied Soft Computing, 49, 137-145.
[14] Pradeep, K. R., & Naveen, N. C. (2016, December). Predictive analysis of diabetes using J48 algorithm of classification techniques. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I) (pp. 347-352). IEEE.
[15] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science, 47, 76-83.
[16] 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.
[17] Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), 127-136.
[18] Wilinska, M. E., Chassin, L. J., Acerini, C. L., Allen, J. M., Dunger, D. B., & Hovorka, R. (2010). Simulation environment to evaluate closed-loop insulin delivery systems in type 1 diabetes. Journal of diabetes science and technology, 4(1), 132-144.
[19] Van Der Heijden, A. A., Ortegon, M. M., Niessen, L. W., Nijpels, G., & Dekker, J. M. (2009). Prediction of coronary heart disease risk in a general, pre-diabetic, and diabetic population during 10 years of follow-up: accuracy of the Framingham, SCORE, and UKPDS risk functions: The Hoorn Study. Diabetes care, 32(11), 2094-2098.
[20] Talmud, P. J., Hingorani, A. D., Cooper, J. A., Marmot, M. G., Brunner, E. J., Kumari, M., ... & Humphries, S. E. (2010). Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. Bmj, 340, b4838.
[21] Pérez-Gandía, C., Facchinetti, A., Sparacino, G., Cobelli, C., Gómez, E. J., Rigla, M., ... & Hernando, M. E. (2010). Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes technology & therapeutics, 12(1), 81-88.

[22] Miyake, K., Yang, W., Hara, K., Yasuda, K., Horikawa, Y., Osawa, H., ... & Ido, K. (2009). Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. Journal of human genetics, 54(4), 236.
[23] Abdul-Ghani, M. A., & DeFronzo, R. A. (2009). Plasma glucose concentration and prediction of future risk of type 2 diabetes. Diabetes Care, 32(suppl 2), S194-S198.
[24] Kilpatrick, E. S., Rigby, A. S., & Atkin, S. L. (2008). Mean blood glucose compared with HbA 1c in the prediction of cardiovascular disease in patients with type 1 diabetes. Diabetologia, 51(2), 365-371.
[25] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017, February). Risk prediction of type II diabetes based on random forest model. In 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) (pp. 382-386). IEEE.
[26] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 1006-1010). IEEE.
[27] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.
[28] Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1584-1589). IEEE.
[29] Ramzan, M. (2016, August). Comparing and evaluating the performance of WEKA classifiers on critical diseases. In 2016 1st India International Conference on Information Processing (IICIP) (pp. 1-4). IEEE.
[30] Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez-Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2017). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. Expert Systems with Applications, 72, 335-343.
[31] Sankaranarayanan, S. (2014, March). Diabetic Prognosis through Data Mining Methods and Techniques. In 2014 International Conference on Intelligent Computing Applications (pp. 162-166). IEEE.
[32] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for type-2 diabetic patients. Expert systems with applications, 37(12), 8102-8108.
[33] kumar Dewangan, A., & Agrawal, P. (2015). Classification of diabetes mellitus using machine learning techniques. International Journal of Engineering and Applied Sciences, 2(5).
[34] Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. In 2014 12th International Conference on Frontiers of Information Technology (pp. 226-231). IEEE.
[35] Mounika, M., Suganya, S. D., Vijayashanthi, B., & Anand, S. K. (2015). Predictive analysis of diabetic treatment using classification algorithm. IJCSIT, 6, 2502-2505.
[36] Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142.
[37] Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208.
[38] Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127). IEEE.
[39] Wang, K. J., Adrian, A. M., Chen, K. H., & Wang, K. M. (2015). An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. Journal of biomedical informatics, 54, 220-229.
[40] Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43), 196-1.
[41] Kang, S., Kang, P., Ko, T., Cho, S., Rhee, S. J., & Yu, K. S. (2015). An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Systems with Applications, 42(9), 4265-4273.
[42] Lukmanto, R. B., & Irwansyah, E. (2015). The early detection of diabetes mellitus (dm) using fuzzy hierarchical model. Procedia Computer Science, 59, 312-319.
[43] Guo, Y., Bai, G., & Hu, Y. (2012, December). Using bayes network for prediction of type-2 diabetes. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 471-472). IEEE.
[44] Li, L. (2014, November). Diagnosis of diabetes using a weight-adjusted voting approach. In 2014 IEEE International Conference on Bioinformatics and Bioengineering (pp. 320-324). IEEE.
[45] Balpande, V. R., & Wajgi, R. D. (2017, February). Prediction and severity estimation of diabetes using data mining technique. In 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 576-580). IEEE.
[46] Hashi, E. K., Zaman, M. S. U., & Hasan, M. R. (2017, February). An expert clinical decision support system to predict disease using classification techniques. In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 396-400). IEEE.
[47] Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE transactions on information technology in biomedicine, 14(4), 1114-1120.
[48] Varma, K. V., Rao, A. A., Lakshmi, T. S. M., & Rao, P. N. (2014). A computational intelligence approach for a better diagnosis of diabetic patients. Computers & Electrical Engineering, 40(5), 1758-1765.
[49] Mekruksavanich, S. (2016, August). Medical expert

system based ontology for diabetes disease diagnosis. In 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 383-389). IEEE.
[50] Nam, J. H., Kim, J., & Choi, H. G. (2015). Developing statistical diagnosis model by discovering principal parameters for Type 2 diabetes mellitus: a case for Korea. Public Health Prev. Med, 1(3), 86-93.
[51] Lee, B. J., Ku, B., Nam, J., Pham, D. D., & Kim, J. Y. (2014). Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE journal of biomedical and health informatics, 18(2), 555-561.
[52] Al Jarullah, A. A. (2011, April). Decision tree discovery for the diagnosis of type II diabetes. In 2011 International conference on innovations in information technology (pp. 303-307). IEEE.
[53] Negi, A., & Jaiswal, V. (2016, December). A first attempt to develop a diabetes prediction method based on different global datasets. In 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 237-241). IEEE.
[54] Pavate, A., & Ansari, N. (2015, September). Risk prediction of disease complications in type 2 diabetes patients using soft computing techniques. In 2015 Fifth International Conference on Advances in Computing and Communications (ICACC) (pp. 371-375). IEEE.
[55] Prajwala, T. R. (2015). A comparative study on decision tree and random forest using R tool. International journal of advanced research in computer and communication engineering, 4(1), 196-199.
[56] Krati Saxena, D., Khan, Z., & Singh, S. (2014). Diagnosis of diabetes mellitus using k nearest neighbor algorithm. International Journal of Computer Science Trends and Technology (IJCST), 2(4), 36-43.
[57] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.
[58] Thirumal, P. C., & Nagarajan, N. (2015). Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. ARPN Journal of Engineering and Applied Science, 10(1), 8-13.
[59] Lavery, L. A., Barnes, S. A., Keith, M. S., Seaman, J. W., & Armstrong, D. G. (2008). Prediction of healing for postoperative diabetic foot wounds based on early wound area progression. Diabetes care, 31(1), 26-29.
[60] Sattar, N., Wannamethee, S. G., & Forouhi, N. G. (2008). Novel biochemical risk factors for type 2 diabetes: pathogenic insights or prediction possibilities?. Diabetologia, 51(6), 926-940.
[61] Van Hoek, M., Dehghan, A., Witteman, J. C., Van Duijn, C. M., Uitterlinden, A. G., Oostra, B. A., ... & Janssens, A. C. J. (2008). Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes, 57(11), 3122-3128.
[62] Lewitt, M. S., Hilding, A., Östenson, C. G., Efendic, S., Brismar, K., & Hall, K. (2008). Insulin-like growth factor-binding protein-1 in the prediction and development of type 2 diabetes in middle-aged Swedish men. Diabetologia, 51(7), 1135.
[63] Larsson, H. E., Hansson, G., Carlsson, A., Cederwall, E., Jonsson, B., Jönsson, B., ... & Ivarsson, S. A. (2008). Children developing type 1 diabetes before 6 years of age have increased linear growth independent of HLA genotypes. Diabetologia, 51(9), 1623.
[64] Chien, K., Cai, T., Hsu, H., Su, T., Chang, W., Chen, M., ... & Hu, F. B. (2009). A prediction model for type 2 diabetes risk among Chinese people. Diabetologia, 52(3), 443.
[65] Grant, R. W., Hivert, M., Pandiscio, J. C., Florez, J. C., Nathan, D. M., & Meigs, J. B. (2009). The clinical application of genetic testing in type 2 diabetes: a patient and physician survey. Diabetologia, 52(11), 2299-2305.
[66] Vergouwe, Y., Soedamah-Muthu, S. S., Zgibor, J., Chaturvedi, N., Forsblom, C., Snell-Bergeon, J. K., ... & Fuller, J. H. (2010). Progression to microalbuminuria in type 1 diabetes: development and validation of a prediction rule. Diabetologia, 53(2), 254-262.
[67] Ekelund, M., Shaat, N., Almgren, P., Groop, L., & Berntorp, K. (2010). Prediction of postpartum diabetes in women with gestational diabetes mellitus. Diabetologia, 53(3), 452-457.
[68] Isomaa, B., Forsén, B., Lahti, K., Holmström, N., Waden, J., Matintupa, O., ... & Tuomi, T. (2010). A family history of diabetes is associated with reduced physical fitness in the Prevalence, Prediction and Prevention of Diabetes (PPP)–Botnia study. Diabetologia, 53(8), 1709-1713.
[69] Maciejewski, M. L., Liu, C. F., & Fihn, S. D. (2009). Performance of comorbidity, risk adjustment, and functional status measures in expenditure prediction for patients with diabetes. Diabetes Care, 32(1), 75-80.
[70] Ziegler, D., Zentai, C. P., Perz, S., Rathmann, W., Haastert, B., Döring, A., & Meisinger, C. (2008). Prediction of mortality using measures of cardiac autonomic dysfunction in the diabetic and nondiabetic population: the MONICA/KORA Augsburg Cohort Study. Diabetes care, 31(3), 556-561.
[71] Sato, K. K., Hayashi, T., Harita, N., Yoneda, T., Nakamura, Y., Endo, G., & Kambe, H. (2009). Combined measurement of fasting plasma glucose and A1C is effective for the prediction of type 2 diabetes: the Kansai Healthcare Study. Diabetes Care, 32(4), 644-646.
[72] Hayes, A. J., Clarke, P. M., Glasziou, P. G., Simes, R. J., Drury, P. L., & Keech, A. C. (2008). Can self-rated health scores be used for risk prediction in patients with type 2 diabetes?. Diabetes care, 31(4), 795-797.
[73] Alehegn, M., Joshi, R., & Mulay, P. (2018). Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm. International Journal of Pure and Applied Mathematics, 118(9), 871-878.
[74] Kleefstra, N., Landman, G. W., Houweling, S. T., Ubink-Veltmaat, L. J., Logtenberg, S. J., Meyboom-de Jong, B., ... & Bilo, H. J. (2008). Prediction of mortality in type 2 diabetes from health-related quality of life (ZODIAC-4). Diabetes Care, 31(5), 932-933.
[75] MacKay, M. F., Haffner, S. M., Wagenknecht, L. E.,

D'Agostino, R. B., & Hanley, A. J. (2009). Prediction of type 2 diabetes using alternate anthropometric measures in a multi-ethnic cohort: the insulin resistance atherosclerosis study. Diabetes care, 32(5), 956-958.

[76] Knip, M., Korhonen, S., Kulmala, P., Veijola, R., Reunanen, A., Raitakari, O. T., ... & Åkerblom, H. K. (2010). Prediction of type 1 diabetes in the general population. Diabetes care, 33(6), 1206-1212.

[77] Kramer, C. K., Von Mühlen, D., Jassal, S. K., & Barrett-Connor, E. (2009). Serum uric acid levels improve prediction of incident type 2 diabetes in individuals with impaired fasting glucose: the Rancho Bernardo Study. Diabetes care, 32(7), 1272-1273.

[78] Ley, S. H., Harris, S. B., Connelly, P. W., Mamakeesick, M., Gittelsohn, J., Hegele, R. A., ... & Hanley, A. J. (2008). Adipokines and incident type 2 diabetes in an Aboriginal Canadian population: the Sandy Lake Health and Diabetes Project. Diabetes Care, 31(7), 1410-1415.

[79] Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making, 10(1), 16.

[80] Bozorgmanesh, M., Hadaegh, F., & Azizi, F. (2010). Diabetes prediction, lipid accumulation product, and adiposity measures; 6-year follow-up: Tehran lipid and glucose study. Lipids in health and disease, 9(1), 45.

[81] Hadaegh, F., Hatami, M., Tohidi, M., Sarbakhsh, P., Saadat, N., & Azizi, F. (2010). Lipid ratios and appropriate cut off values for prediction of diabetes: a cohort of Iranian men and women. Lipids in health and disease, 9(1), 85.

[82] Stern, M., Williams, K., Eddy, D., & Kahn, R. (2008). Validation of prediction of diabetes by the Archimedes model and comparison with other predicting models. Diabetes care, 31(8), 1670-1671.

[83] Ford, E. S., Li, C., & Sattar, N. (2008). Metabolic syndrome and incident diabetes: current state of the evidence. Diabetes care, 31(9), 1898-1904.

[84] Olmsted, M. P., Colton, P. A., Daneman, D., Rydall, A. C., & Rodin, G. M. (2008). Prediction of the onset of disturbed eating behavior in adolescent girls with type 1 diabetes. Diabetes Care, 31(10), 1978-1982.

[85] Cederholm, J., Eeg-Olofsson, K., Eliasson, B., Zethelius, B., Nilsson, P. M., & Gudbjörnsdottir, S. (2008). Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. Diabetes care, 31(10), 2038-2043.

[86] Schulze, M. B., Weikert, C., Pischon, T., Bergmann, M. M., Al-Hasani, H., Schleicher, E., ... & Joost, H. G. (2009). Use of multiple metabolic and genetic markers to improve the prediction of type 2 diabetes: the EPIC-Potsdam Study. Diabetes care, 32(11), 2116-2119.

[87] Siljander, H. T., Simell, S., Hekkala, A., Lähde, J., Simell, T., Vähäsalo, P., ... & Knip, M. (2009). Predictive characteristics of diabetes-associated autoantibodies among children with HLA-conferred disease susceptibility in the general population. Diabetes, 58(12), 2835-2842.

[88] Savvidou, M., Nelson, S. M., Makgoba, M., Messow, C. M., Sattar, N., & Nicolaides, K. (2010). First-trimester prediction of gestational diabetes mellitus: examining the potential of combining maternal characteristics and laboratory measures. Diabetes, 59(12), 3017-3022.

[89] Gillis, R., Palerm, C. C., Zisser, H., Jovanovic, L., Seborg, D. E., & Doyle III, F. J. (2007). Glucose estimation and prediction through meal responses using ambulatory subject data for advisory mode model predictive control.

[90] Finan, D. A., Doyle III, F. J., Palerm, C. C., Bevier, W. C., Zisser, H. C., Jovanovič, L., & Seborg, D. E. (2009). Experimental evaluation of a recursive model identification technique for type 1 diabetes. Journal of diabetes science and technology, 3(5), 1192-1202.

[91] Hippisley-Cox, J., Coupland, C., Robson, J., Sheikh, A., & Brindle, P. (2009). Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. Bmj, 338, b880.

[92] Clayton, D. G. (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS genetics, 5(7), e1000540.

[93] Abdul-Ghani, M. A., Lyssenko, V., Tuomi, T., DeFronzo, R. A., & Groop, L. (2009). Fasting versus postload plasma glucose concentration and the risk for future type 2 diabetes: results from the Botnia Study. Diabetes care, 32(2), 281-286.

[94] Abu-Naser, S. S., Kashkash, K. A., & Fayyad, M. (2010). Developing an expert system for plant disease diagnosis. Journal of Artificial Intelligence, 3(4), 269-276.

[95] Er, O., Yumusak, N., & Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. Expert Systems with Applications, 37(12), 7648-7655.

[96] Dogantekin, E., Dogantekin, A., Avci, D., & Avci, L. (2010). An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digital Signal Processing, 20(4), 1248-1255.

[97] Törn, C., Mueller, P. W., Schlosser, M., Bonifacio, E., & Bingley, P. J. (2008). Diabetes Antibody Standardization Program: evaluation of assays for autoantibodies to glutamic acid decarboxylase and islet antigen-2. Diabetologia, 51(5), 846-852.