International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach

Minyechil Alehegn[1], Rahul Joshi[2]

[1]Dept. of Computer science and engineering, Symbiosis Institute of Technology, Pune - 412115, Maharashtra, India
[2] Assistant Professor Dept. of Computer science and engineering, Symbiosis Institute of Technology, Pune - 412115, Maharashtra, India

---

Abstract - *Machine learning techniques (MLT) are used to predict the medical datasets at an early stage of safe human life. A huge medical datasets are accessible in different data repositories which used to in the real world application. Now a day Machine learning (ML) has the ability to answer questions. One of the missions is a prediction on disease data. Currently Diabetes Diseases (DD) are among the leading cause of death in the world. To group and predict symptoms in medical data, various data mining techniques were used by different researchers in different time. A total of 768 instances, data set from PIDD (Pima Indian Diabetes Data Set). In this system the most known predictive algorithms apply KNN, Naïve Bayes, Random forest, and J48. By using these algorithms make an ensemble hybrid model by combining individual techniques/methods into one in order to increase the performance and accuracy.*

Key words: Ensemble, Diabetes, classification, Machine learning, Data mining, KNN, Naïve Bayes, Random Forest, J48.

1. INTRODUCTION

Diabetes diseases commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic diseases in which the person has blood sugar, either insulin production inefficient, or because of the body cell do not return correctly to insulin, or by both reason. The day is now to prevent and diagnose diabetes in the early stages.

According to the WHO (world health organization) report in Nov 14, 2016 in the world diabetes day "*Eye on diabetes*" reported 422 million adults are with diabetes, 1.6 million deaths, as the report indicates it is not difficult to guess how much diabetes is very serious and chronic.

In 2014, 8.5% of adults whose ages are 18 and older than 18 had diabetes. In 2012 HBG (high blood glucose was the cause of 2.2 million people deaths [53]

Diabetes diseases damage different parts of the human body from those parts some of them are: eyes, kidney, heart, and nerves. *Williams's textbook of endocrinology was* predictable that in 2013 more than 382 million population in the world or all over the world were with diabetes or had diabetes. There are so many peoples are died every year by diabetes disease (DD) both in poor and rich countries in the world.

According to the centers for disease control and prevention (CDCP) they give information for the duration of 9 ensuing years that is between 2001 and 2009 type II diabetes increased 23% in the United States (US). There are different countries, organization, and different health sectors worry about this chronic disease control and prevent before the person death.

Diabetes. Most in the current time diabetes is grouped into two types of diabetes, type I and Type II diabetes. Type I diabetes this type of diabetes in heath language or in doctors' language this type of diabetes also called Insulin dependent diabetes illness. Here the human body does not produce enough insulin. 10 % of diabetes caused by this type of diabetes.

Type II diabetes this type of diabetes. According to CDA (Canadian Diabetes Association) during 10 years, between 2010 and 2020, expected to increase from 2.5 million to 3.7 million. Therefore, as the above mentioned Diabetes diseases needs early preventation and diagnosis to safe human life from early death .By considering how much this disseises is very series and leading one in the world. Moloud et al. [2] Algorithms which are used in machine learning have various power in both classification and predicting.

Abdullah et al. [40] Data mining methods support health care researchers to retrieve novel knowledge from large health data. With the development of Information Technology, Data mining offers appreciated advantage in diabetes research,

which leads to expand or improve health care distribution, increase support for decision –making and improve disease supervision.

Saba et al. [12] no single technique gives highest accuracy or accuracy for all diseases, whereas one classifier provides or shows better performance in a given dataset, another method or approach outdoes the others for other diseases. The new study or the proposed study concentrates on a novel combination of different classifiers for diabetes disease (DD) classification and prediction, thus overcoming the problem of individual or single classifiers.

This study follows different machine learning algorithms to predict diabetes disease at an early stage.Such as, KNN, Naïve Bayes, Random Forest, and J48 to predict this chronic disease at an early stage for safe human life.

2. RELATED WORK

Song et al. [**8**] Describe and explain different classification Algorithms using different parameters such as Glucose, Blood Pressure, Skin Thickness, insulin, BMI, Diabetes Pedigree, and age. The researches were not included pregnancy parameter to predict diabetes disease (DD). In this research, the researchers were using only small sample data for prediction of Diabetes. The algorithms were used by this paper were five different algorithms GMM, ANN, SVM, EM, and Logistic regression. Finally. The researchers conclude that ANN (Artificial Neural Network) was providing High accuracy for prediction of Diabetes.

Loannis et al.[7] machine learning algorithms are very important to predict different medical data sets including diabetes diseases dataset(DDD).in this study they use support vector machines(SVM) ,Logistic Regression ,and Naïve Bayes using 10 fold cross validation to predict different/varies medical datasets including diabetes dataset(DD) .the researchers' was compare the accuracy and the performance of the algorithm based on their result and the researchers conclude that SVM(support Vector Machine) algorithm provides best accuracy than the other algorithm which are mentioned on the above . The researchers were use those machine learning algorithm on a small sample of data.in this study factors for accuracy were identified such factors are Data origin, Kind, and dimensionality.

Nilashi et al. [9] .CART (classification and Regression Tree) was used for generating fuzzy rule. Clustering algorithm also was used (principal component Analysis (PCA) and Expectation maximization (EM) for pre-processing and noise removing before applying the rule. Different medical dataset (MD) was used such as breast cancer, Heart, and Diabetes Develop decision support for different diseases including diabetes. The result was CART (Classification and Regression tree) with noise removal can provide effective and better in health/diseases prediction and it is possible to safe human life from early death.

Yunsheng et al. [1] this study was the new approach that used KNN algorithm by removing the outlier/OOB(out of bag) using DISKR(decrease the size of the training set for K-nearest neighbour .and also in this study the storage space was minimized. There for ,the space complexity is become less and efficient .after removing a parameters or instances which have less effect or factor the researchers got better accuracy .

Francesco et al.[4]feature selection is one of the most important step to increase the accuracy.Hoeffding Tree(HT) ,multi-layer perceptron(MP),Jrip,BayeNet,RF(random forest),and Decision Tree machine learning Algorithms were used for prediction .From different feature selection algorithm in this study they were use best first and greedy stepwise feature selection algorithm for feature selection purpose . The researchers conclude that Hoeffding Tree (HT) provides high accuracy.

Pradeep et al.[29]in this study the researchers concentrate on different datasets including Diabetes Dataset(DD).The researcher were investigate and construct the models that are universally good and capability for varies/different medical datasets (MDs).the classification algorithm did not evaluate using Cross validation evaluation method .

ANN,KNN,Navie Bayes,J48,ZeroR,Cv Parameter selection, filtered classifier ,and simple cart were some of the algorithm used in this study. From those algorithm Naïve Bayes provide better accuracy in diabetes dataset (DD) in this study. The two algorithm KNN and ANN provide high accuracy in other datasets on this study.

Sajida et al.[16]by using CPCSSN(Canadian primary care sentinel surveillance Network) dataset and three machine learning methods to predict the diabetes Disses (DD) in early stage to safe human life at from early death .on this study Bagging ,Adaboost,and decision tree(J48) were used to predict the diabetes and the researcher was compare the result of those methods and concluded that Adaboost method was provide effective and better accuracy than the other methods in weka data mining tools

Kamadi et al. [17] classification problems were identified in this study.one of the most problem in classification is data reduction .it has a vital role in prediction accuracy .to get better and efficient accuracy the data should be reduced as the researchers studied here. On this study PCA (principal component Analysis) for data pre-processing including data reduction for better accuracy. For prediction modified decision tree (DT) and Fuzzy were used for prediction purpose .finally it was concluded as to get better result the dataset should be reduced.

Pradeep & Dr.Naveen [15] in this study the performance of machine learning techniques were compared and measured based on their accuracy. The accuracy of the technique is vary from before pre-processing and after pre-processing as they identified on this study. This indicates the in the prediction of diseases the pre-processing of data set has its own impact on on the performance and accuracy of the prediction

Decision tree technique provide better accuracy in this study before pre-processing to predict diabetes diseases. Random forest and support vector machine provides better prediction after pre-processing in this study using diabetes data set.

Santhanam and Padmavathi [21]K-means and Genetic algorithm used in this study for Dimension reduction in order to get better performance. The integration of support vector machine for prediction technique was used and provide better accuracy in small sample diabetes data set by selecting only five factors or parameters. 10 cross validation on this study used as evaluation method.finaly reduced data set provide better performance than large dataset.

Xue-Hui Meng et al. [42] in this study the researchers were use different data mining techniques to predict the diabetic diseases using real world data sets by collecting information by distributed questioner .in this study SPSS and weka tools were used for data analysis and prediction respectively .in this study the researchers compare three techniques ANN, Logistic regression, and j48 .finally it was concluded as j48 machine learning technique provide efficient and better accuracy.

Abdullah et al. [40] Oracle Data miner and Oracle Database 10g used for Analysis and storage respectively .the parameters or factors were identified in this study .the target variables were identified based on their percentage .this study concentrated on the treatment of the patient .the patient divided into two categories old and young based on their age and predict their treatment .for both young and old diet controle indicates high percentage on this study. The treatment predictive percentage done by support vector machine.

3. METHODOLOGY

In diabetic disease there were different research were done .previously there were many researchers did different researches in health care centres. From those researchers money of them also did on diabetes disease as it was series issues in the old aged research done only on the health centres not in the computerised like machine learning approach .it is also true now a day summary of common or major findings are given as follow in the form of table.

Table I: Summary of major findings or discoveries of diabetes prediction methodologies

Sn	Authors	Methodologies	Findings
1	Weifeng Xu et al.[6]	Naïv Bayes Random forest ID3 Adaboost	Random forest was better than other. ID3 was provided less accuracy than others.
2	Messan et al.[8]	ANN,GMM,SVM, Logistic Regression, and ELM	ANN was best accuracy relative to others.
3	Loannis et al.[7]	Logistic regression Naïve Bayes Svm	In this study svm with accuracy of 84% with 10 fold cross validation
4	Mehrbakhsh et al.[9]	CART,clustering Algorithm(PCA and EM)	Some fuzzy rules were generated by CART. Fuzzy rule based ,and CART by removing noise was effective in prediction purpose
5	Tao et al.[3]	KNN,Naïve Bayes, Random Forest, decision tree, svm, and logistic regression ,	Filtering criteria was improved. The accuracy of recall was better in this study.

6	Yunsheng et al.[1]	KNN,DISKR	In this study the storage space was reduced, an instance which have less factor was eliminated. Removing of outlier increase accuracy.
7	Francesco et al.[4]	Hoeffding,j48,multilayer perceptron,Jrip,Bayenet, ,Best first ,Greedy stepwise , and Random Forest	In this study feature selection was the main targeted. 10 fold cross validation was used for splitting mechanism Hoeffding was provide better accuracy by integrating with searching algorithm with 77.5% than others.
8	Swarupa et al.[14]	Naïve Bayes ANN,KNN,J48,zeroR,cv parameter selection ,simple cart, and Filtered classifier	In this paper different dataset applied including diabetes In this study any cross validation technique was not applied. Naive Bayes was provide high accuracy with the accuracy value of 77.01%.
9	Sajida et al.[16]	Bagging,Adaboost,and j48	In this study the researchers have got Adaboost as the better accuracy relative to others.
10	Munaza Ramzan[19]	Naïve Bayes,Random Forest,and J48	Random forest was provided better accuracy than J48 and Naïve Bayes in 10 cross validation splitting method.
11	Kamadi et al.[17]	Modified fuzzy and PCA	Data reduction was applied in this study.to got the better accuracy the data should reduce
12	Pradeep & Dr.Naveen [15]	J48	It was one of most popular and noted as better accuracy in this study .feature selection was applied.
13	Ramiro et al.[5]	Fuzzy rule	In this study recommended system was developed, it was help to reduce the wrong treatment.
14	Pradeep et al.[29]	J48,KNN,Random Forest ,and SVM	The algorithm were compared and j48 was provided better accuracy by providing 73.82% than others before pre-processing .KNN and RF were provided good accuracy after pre-processing .
15	Santhanam and Padmavathi[21]	K-means,Genetic Algorithm ,and SVM	New integrated system clustering and classification algorithm and shown high accuracy.
16	Sankarana &Dr Pramananda[37]	Association rule using apriori and FP growth.	Fast and better clinical decision making helps for preventive and suggestive medicine Fp growth was more advantages over apriori
17	Xue-Hui Men et al.[42]	J48,Logistic Regression, and KNN	There were comparison between the algorithms performance and j48 shown high accuracy with 78.27%.
18	Abdullah et al.[40]	SVM	This study concentrated on the effective treatment prediction.
19	Patil et al.[47]	HPM	It was efficient and better accuracy by providing 92.38%
20	Saba et al.[12]	HMV,NB,Adaboost,RF SVM,KNN,and LR	Was concentrated on different diseases including diabetes .HMV were provided high accuracy than others with the accuracy of 78.085
21	Amit and Pragati [30]	C4.5,RF,MLP,and Bayes Net	Hybrid model was applied. From the algorithm the hybrid of MLP+BayesNet provided high accuracy of 81.89%
22	Saba et al.[35]	ID3,C4.5 ,Bagging ,and	Bagging was shown high accuracy than other

		CART	techniques.
23	Mounika et al.[32]	ZeroR,oneR,and Naïve Bayes	Effective treatment in young and old patient were studied. Naive Bayes was better performance than others
24	Nongyao and Rungruttikarn[33]	LR, NB, ANN, Bagging, Boosting, and Decision tree.	Hybrid concept was apply by using bagging or boosting .RF provided high accuracy of 85.558
25	Dr Saravana et al.[31]	Predictive analysis algorithm in Hadoop	Concentrated on treatment in health care industry using big data analysis. The result shown that proper treatment with low cost
26	Veena and Anjali[23]	SVM,NB,Decision Stump, and decision tree	Hybridization concept was done on this study using the base classifier with bagging .Decision stump with provided better accuracy of 80.72%
27	Kung et al.[34]	Novel EM method ,oposit sign test, and KNN	New and effective feature selection mechanism done on this study by hybridizing EM and KNN.
28	Saravananatha n and velmurugan[18]	J48,CART,SVM,and KNN	In this study j48, cart, svm and knn was applied and provide 67.15%, 62.28, 65.04 and 53.39 respectively.
29	Seokho et al.[28]	SVM,E^2_SVM	This study was concentrated on drug failure prediction .this study was good and ensemble approach. E^2_SVM was shown better accuracy than single Svm with accuracy of 80 %.
30	Rian and Irwansyah[27]	Fuzzy rule	Rules were generated in this study that were helps early detection.
31	Yang et al.[43]	Naïve Bayes, Bayes network.	Bays network was provided high accuracy of 72.3%
32	Lin[39]	SVM,ANN,Naïve Bayes,	Weighted Adjusted based study. The majority voting was applied in this study. The combination of the classifier were provide better accuracy than the single one
33	Vrushali and Rakhi[10]	CLAT	Prediction and severity estimation of diabetes in different bodies were done.
34	Emrana et al.[11]	C4.5 and KNN	In this study c4.5 and knn technique were provided with accuracy of 90.43 and 76.96 % respectively
35	Nahla et al[46]	SVM with rule extraction with SQRex-SVM	In this stud the combined model provided high accuracy.
36	Kamadi et al.[38]	Decision Tree, Gini index, Gaussian fuzzy function	Decision tree model was provided better accuracy
37	Sakorn[13]	Expert system with fuzzy rule	In this paper expert system for treatment was done.
38	Ayush and Divya[24]	CART	This algorithm was provided accuracy of 75%
39	Jae et al.[26]	Wrapper and linear forward selection	The computation time was reduced in this study.
40	Bum et al.[36]	Logistic regression and Naïve Bayes, Anthropometry	It was focused on prediction of Fasting Glucose Level. Here the better accuracy was 74.1%
41	Asma [45]	Decision tree	Decision tree was provided good result with the accuracy of 78.1768%
42	Anjli and Varun[20]	SVM	In this study feature selection was done using wrapper and ranker .SVM shown accuracy of 72% with ranker feature

			selection. Percentage split was applied.
43	Aruna and Nazneen[25]	KNN, fuzzy rule, and GA	In this study association between KNN and GA were done. Some rule was generated.
44	Prajwala[22]	RF and DT	RF was provided good accuracy than DT .execution time for RF was more than DT in this study.
45	Emirhan et al.[44]	ANFIS, Rough Set	In this work ANFIS was provide better result than Rough Set .
46	Krati et al.[48]	KNN	was gotten the accuracy of 70% in data tes1 and 57% in data test2 respectively
47	Anuja and Chitra[41]	SVM	Svm was provided the accuracy of 78%
48	Thirumal et al.[49]	Naïve Bayes,SVM,KNN,C4.5	In this study c4.5 was shown better than other with accuracy of 78.2552%

3.1 Data pre-processing Methods

The data that we used must be wisely composed, joined/integrated and ready for analysis [42].

The dataset used in this study obtained from public UCI repository PIDD (Pima Indian Diabetes Database) which is available online .we will use this online available dataset for analysis and prediction of diabetes diseases. This diabetes dataset consists 768 records and 8 attributes with one target class.in this study Weka 3.8.1 and java using netbean 8.2 use for analysis, classification, and prediction. And also, Ensemble hybrid model with base learner for prediction is include.

3.2 Classification and prediction Methods

In this study, the following parameters are used as input pregnancies, Glucose, Blood Pressure, skin thickness, insulin, BMI, Diabetes pedigree Function, and Age. There are a number of machine learning and statistical techniques that can used to predict diabetes diseases. Based on the extent literature, we settled on employing four most known machine learning algorithm (Random Forest (RF), KNN, Naïve Bayes, and J48) classification algorithm and ensemble/combined them in to one using base learner. The following section describes these Classification techniques and their unique requirements used in this research study.

Random forest (RF)

RF is one of the popular and adaptable algorithm used in ensemble technique .it is the best and popular machine learning algorithm in the concept of hybrid model for the improvement performance and prediction accuracy.RF is easy to handle large data and high dimensionality. The samples are selected arbitrarily.

KNN

K-Nearest Neighbour algorithm is one of the classification algorithm .it is the simplest and easy than other data mining techniques .this technique classifies new belongings based on similarity measure [18].the value of k always assign positive integer number .In this algorithm the training data are stored .based on the neighbours or nearest prediction of test data is complete

Step/phase I. Determine k which is the number of nearby neighbours.

Step II/phase. Estimate distance between the instance and training samples.

Step/phase III: The remoteness of the training samples are sorted and the closest neighbour based on the minimum the distance is determined in this step.

Step/phase IV: in this step we get all the classes of all the training data

Step/phase V: use the majority of the class of closest neighbours as the prediction value of the query instance.

Naïve Bayes (NB)

Naïve Bayes (NB) is one of the most popular and suitable when the imputes is large .this machine learning method

or technique need a short time complexity or computational time. NB computes based on possibility by using Bayes formula [19].

J48

It is an improvement of ID3 classification algorithm .j48 has the ability of select a specific parameters or instances and lost attribute. This type of classification algorithm has the ability to support continuous as well as categorical instances in the process of tree construction rules which are constructed by this algorithm are easy and simple to understand [47].

Hybrid model

In prediction individual classification algorithms are not provided result so, it is better to make the result of those individual classifier in to one by combining the prediction of individual classifier.an ensemble approach the problem or limitation of distinct classifiers to increases the accuracy by combining in to one. [12, 47].hybrid model provides best performance and accuracy than the single one that is the reason why this method chosen in this study.

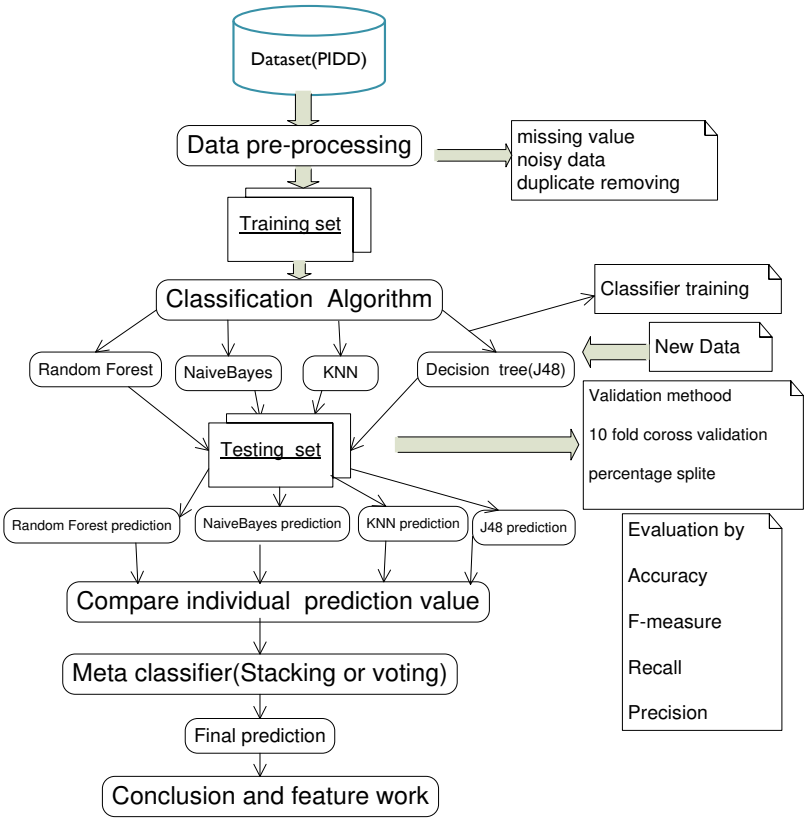

Fig1:- Detail Architecture of work flow

OBJECTIVE OF STUDY

The main goal of this analysis study is predict the diabetes disease and compare the algorithm which algorithm provide high accuracy .finally select the best algorithm to predict the diabetes disease at early stage. Examine how patients' characteristics as well as measurements disturb diabetes cases.

4. CONCLUSION

Various data mining techniques and its application were studied or reviewed .application of machine learning algorithm were applied in different medical data sets Machine learning methods have different power in different data set. Single algorithm provided less accuracy than ensemble one.in most study decision tree provided high accuracy.in this study hybrid system Weka and java are the tools to predict diabetes dataset.

ACKNOWLEDGEMENT

First of all I would like to thank the Almighty God and his mother merry Mariam for their unending blessings. I would like to express my great attitude to my Research guide professor Rahul Joshi deep regard for his model guidance, feedback, suggestion and constant encouragement. And also I would like to express attitude to the reviewer of the paper and their value able suggestion. Finally I would like to thank to my friends and parents for their support.

REFERENCES

[1] Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.

[2] 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.

[3] 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.

[4] Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. Procedia Computer Science, 112(C), 2519-2528.

[5] 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.

[6] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017, February). 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.

[7] 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.

[8] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2017, June). Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2017 2nd International Conference on (pp. 1006-1010). IEEE.

[9] Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering.

[10] Balpande, V. R., & Wajgi, R. D. (2017, February). Prediction and severity estimation of diabetes using data mining technique. In Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on (pp. 576-580). IEEE.

[11] 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 Electrical, Computer and Communication Engineering (ECCE), International Conference on(pp. 396-400). IEEE.

[12] 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.

[13] Mekruksavanich, S. (2016, August). 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.

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

[15] Pradeep, K. R., & Naveen, N. C. (2016, December). 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.

[16] 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.

[17] 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.

[18] Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43).

[19] Ramzan, M. (2016, August). 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.

[20] Negi, A., & Jaiswal, V. (2016, December). 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.

[21] 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.

[22] 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, 196-1.

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

[24] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In Next Generation Computing Technologies (NGCT), 2015 1st International Conference on (pp. 673-676). IEEE.

[25] Pavate, A., & Ansari, N. (2015, September). 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.

[26] 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.

[27] Lukmanto, R. B., & Irwansyah, E. (2015). The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model. Procedia Computer Science, 59, 312-319.

[28] 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.

[29] Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.

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

[31] Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208.

[32] Mounika, M., Suganya, S. D., Vijayashanthi, B., & Anand, S. K. (2015). Predictive analysis of diabetic treatment using classification algorithm. IJCSIT, 6, 2502-2505.

[33] Nai-arun, N., & Moungmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. Procedia Computer Science, 69, 132-142.

[34] 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.

[35] 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 Frontiers of Information Technology (FIT), 2014 12th International Conference on (pp. 226-231). IEEE.

[36] 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.

[37] Sankaranarayanan, S. (2014, March). Diabetic prognosis through Data Mining Methods and Techniques. In Intelligent Computing Applications (ICICA), 2014 International Conference on (pp. 162-166). IEEE.

[38] 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.

[39] Li, L. (2014, November). Diagnosis of Diabetes Using a Weight-Adjusted Voting Approach. In Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on (pp. 320-324). IEEE.

[40] 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.

[41] 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.

[42] 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.

[43] Guo, Y., Bai, G., & Hu, Y. (2012, December). Using bayes network for prediction of type-2 diabetes. In Internet Technology And Secured Transactions, 2012 International Conference for (pp. 471-472). IEEE.

[44] Yıldırım, E. G., Karahoca, A., & Uçar, T. (2011). Dosage planning for diabetes patients using data mining methods. Procedia Computer Science, 3, 1374-1380.

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

[46] 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.

[47] 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.//19

[48] Krati Saxena, D., Khan, Z., & Singh, S.(2014) Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm.

[49] 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).

[50] Chandna, D. (2014). Diagnosis of heart disease using data mining algorithm. (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2), 1678-1680.

[51] Khemphila, A., & Boonjing, V. (2010, October). Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. In Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on (pp. 193-198). IEEE.

[52] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

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

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

BIOGRAPHIES

Rahul Joshi is presently pursuing PhD at Symbiosis Institute of Technology, India till now. He is received M.Tech from IIT, Mumba i, India. He worked at Symbiosis Institute of Technology as Assistant Professor .He Worked as a Software Developer in ASCIPL, Mumbai from June 2010 to May 2011. His research interest include Machine learning, Data mining, Networking, NLP, Big Data, and Artificial Intelligence.

Minyechil Alehegn is currently M.Tech Candidate in the department of computer Science and Engineering Symbiosis Institute of Technology, India. He is Received his B.SC. Degree in Information Technology from Wollega University, Ethiopia .He worked at Mizan Tepi University from 2014 to 2015 as lecturer. His research interest include Machine learning, Data mining, NLP, and Artificial Intelligence.