Forecast Regression analysis for Diabetes Growth: An inclusive data mining approach

M.N. Sohail, Ren Jiadong, M.M. Uba, M. Irshad, Musavir Bilal, Usman Akbar, Tahir Rizwan

Abstract— **From the past decade, data mining is involved in many fields as well the tremendous work has carried out on the medical sectors to diagnose different diseases like heart strokes, kidney and diabetes. Many applications have been introduced in this manifesto, which can be classified in to two sets of branches: the policy development and the decision support analysis. Still there is a lack of work in decision-making section. Our paper has aim towards the set of decision-making through classification analysis forecast. Data mining approach has work on this manifesto for us to analyze and forecast the analysis for the better clinical decision-making process by machine learning classification and logistic regression modeling has used to forecast the outcome of dataset through classification. Our proposed aim compared the classification analysis of previous researchers work by weka classification of experimenter part and auto-weka guide for better predictions. Our path approaches the best analysis in prediction through forecast and present the comparison of different classification techniques applied on the data set.**

Index Terms— **algorithms, classification, diabetes mellitus, forecast, healthcare data mining, KDD, and regression**

I. Introduction

According to the doctors **[1]**, Diabetes is mostly referred as diabetes mellitus; which is spreading vastly around the globe. Around 425 million people had diabetes in the world by the survey report of International Federation in ‘2015’ and this ratio can be expected to the 700 million by 2040.

The name diabetes discovered by the Greek physician during the second century A.D. **[2]**’ which has meaning of “siphons”. He observed that ants would be attracted to some people’s urine and described diabetes from the patients’ drinking to much water (polyuria) like a siphon. Later English has adopted the word Diabetes. Diabetes is commonly known as the group of metabolic diseases **[3]**, which indicates the high blood pressure and sugar level for the prolonged period. The most common symptoms are frequent urination, increased in hunger and also in thirst **[4]**. These symptoms can cause so many other harshly problems, if it left untreated on early stages. The most commonly known complications occur in the body is hyperosmolar, diabetic ketoacidosis, hyperglycemic and also death of human body **[5]**. Diabetes also cause the long term complications known as cardiovascular diseases, heart strokes, kidney failure, chronic ulcers, vision blindness, damage of eyes and many more,’ which we have illustrated in the section of dataset attributes as an attribute collections of real-life data for this research.

1. In Type 1 diabetes **[6]**, body refuses to produce insulin. In this case, the immune system of the patient body attacks and destroys the cells in pancreases, which produce insulin’s. Utmost the 10% of all cases registered in hospitals refers to Type 1. Which has been diagnosed in mostly children’s and young adults. People with this type of diabetes need to take the insulin injections every day to keep alive the body cells.
2. In Type 2 diabetes **[7]**, the body does not produce enough amount of insulin for functioning the cells and sometimes body does not take insulin’s injection well. This type can be developing in any age of human body but according to our research it occurs mostly in the middle age and older age. About 90% of cases registered in hospital around the world are Type 2.
3. Gestational diabetes occurs in women’s **[8]**, mostly the pregnant females has this case but it disappears’ when the baby born. According to the doctors, the female’s who has Gestational diabetes seems to have the Type 2 later in the life.

Most of the research has been carried out on Kidney issues **[5]** or heart problems **[9]** but the research survey of National Institute of Diabetes and Digestive and Kidney Diseases **[10]** in 2015 says, 30 million people of United States have diabetes’, which calculated as the 9.5% of the total population. From 1 in 4 of them does not know that they have disease. Diabetes affects 1 in 4 people’s over the age of 60’s and about 95% cases been noticed of Type 2 in adults. Mostly the people of 40 years of age have Type 2 cause of family history or the over weight problems **[11]**. Some of them are physically inactive on way of exercise or some have high blood pressure and some of them had diabetes in early ages or have Gestational on the stage of pregnancy. Medical specialists can determine the diabetes by three different tests

Manuscript received September 2018.

M.N. Sohail. *Department of Information Sciences and Technology, Yanshan University, China. (E-mail: mn.sohail@stumail.ysu.edu.cn).*

Ren Jiadong. *Department of Information Sciences and Technology, Yanshan University, China. (E-mail: jdren@ysu.edu.cn).*

M.M. Uba. *Department of Information Sciences and Technology, Yanshan University, China. (E-mail: musaubamuhammad@gmail.com).*

M. Irshad. *Department of Information Sciences and Technology, Yanshan University, China. (E-mail: ibrahim@stumail.ysu.edu.cn)*

Musavir Bilal. *Department of Information Sciences and Technology, Yanshan University, China. (E-mail: musavirbilal@stumail.ysu.edu.cn).*

Usman Akbar. *Department of Economics and Management, Yanshan University, China. (E-mail: usman.akbar@stumail.ysu.edu.cn).*

Tahir Rizwan. *Department of Controls and Engineering, Shanghai Jiao Tong University, China. (E-mail: tahirrizwan@sjtu.edu.cn).*

on patients **[12]**:

1. A1C:
   i) If test results are 6.5%, means patient has diabetes.
   ii) If test results are between 5.7 to 5.99%, means patient has diabetes before.
   iii) If test results are <5.7%, means patient has no diabetes.
2. Fasting plasma glucose test:
   i) If FPG is 126 mg/dl, means patient has diabetes.
   ii) If FPG is between 100 to 125.99 mg/dl, means patient has diabetes before.
   iii) If FPG is <100 mg/dl, means patient is normal.
3. Oral glucose tolerance test:
   i) If OGTT is 200 mg/dl, means patient has diabetes.
   ii) If OGTT is between 140 to 199.9 mg/dl, means patient has diabetes before.
   iii) If OGTT is <140 mg/dl, means patient is normal.

The remaining part of the paper is divided in to five sections. Section 2 will talks about the related work carried out by different researchers from past decade. Section 3 will shows the data processing methodology. Section 4 will defend the paper by illustrating the driven results achieved in experiment and lastly we will conclude the paper with references.

II. Related work

Data mining is the operation of extracting data from many sides to figure out the patterns that produce useful results in the business growth **[2], [11]**. It's playing a tremendous workout in the field of technology to sort out the data from various datasets of different domains. In example of healthcare sectors, data mining is extracting the hidden patterns in the data for the treatment, predictions and diagnosis of various harmful diseases like cancer and HIV. Because of the fruitful results of data mining, the industries are improving their quality of services **[3]**. It is useful in medical applications like medicines, therapeutic experiments, operation measures and medical catalogs **[13]**. Apriori and FpGrowth are extensively used in regular pattern mining procedures **[14]**. The two common algorithms were being reviewed in **[2], [10], [11], [15]–[17]**. These are consumed in therapeutic data mining. Birth consequences were driven in **[18]**. Genetic disease were measured by data mining techniques in **[19]**. C. Kruse et.al has worked on medical data mining with the Neural Networks for the data visual images **[20]**. In early 2000, data mining terminology and techniques were used to mine the medical hypermedia classification gist **[21]**. Missing values theory in medical datasets was acknowledged in **[22]**. More over Morgan et.al **[23]** has analyzed and discovered the figurative rule mining worktable for spawning initial rule sets and Hoe has controlled in the same projects to discover the rule sets. He has examined the rule set as consequence of intelligence's mining for edifice rule based proficient techniques. Wolfsdorf et.al **[24]** has investigated the linear genetic platform and Neural Networks for the medicinal data mining. F. Jia **[25]** has suggested the association rule algorithm for mining the medical images and produced the results that can discover the habitually arising stuffs with-in the data sets. Jothi et.al stated in article **[19]** that Olukandu has worked on the classification system, which livelihoods the Bayesian Ying Yang principle and successfully smeared it to predict the liver disorder and other liver diseases. Regarding mining the medical data of Geno, Morgan et.al **[23]** has introduced a structure, which is centered by heterogeneous and grid distribution.

For the medical image analysis, Fallah et al. **[26]** has discovered a call tree data-mining algorithm and also Maina et al. **[27]** has used Fuzzy Cluster algorithm to mine the medical images, decision tree was been used in their study to classify the irregular and outdated cases. Which was been studied on the lung cancer disease diagnosis and X-ray images. One more outlier prediction method was been discovered by Jia et al. **[25]** to classify the medical data. Cardio Vascular diseases has diagnosed with the help of classification algorithm by Ahmed et al. **[28]**. He aims to the two-chin detection method, which includes impulsive features choice and proficient verdict. Mirza et al. **[29]** has worked on tele- medicines and introduced web based data mining for it. Xiong et al. **[30]** studied on patient data sets and discussed a method to amalgamate the rule mining methods and classification methods. In his analysis, he recycled the Swarm Optimization method. The moral result shows that their research is valuable in Brain disease prediction and diagnosis. Muhammad et al. **[15]** urbanized an association rule initiate that is based on the earlier Breast cancer patients. The rule initiate is active during "clinical Trial Assignment Professional System". Patil et al. **[31]** applied Bayesian algorithm for spotting of confusion Coronary Cardiopathy. Time Annotated Sequence Rule was been studied by Chalew et al. **[32]** to discover the temporal measurements from medical datasets. The discovered patterns help in improvement of identification by attributes interactions in the entire time sphere.

Mostafa et al. **[33]** has worked on data mining technique in prediction of survival CHD patients. He did combine the three prediction classification tools as: Support Vector Machine, Artificial Neural Networks and Decision Tree (C4.5, ID3, CART and C5) in prediction of CHID patients. Evidentially their result shows that the techniques and methods they used are a long way efficient in analyzing and handling the different disease data in to factions to support the chronological covariance of Choline esterase, albumin and living affluence accordingly.

Beside all Larsson et al. **[6]** practically applied Apriori method to mine the medical data. He isolated the frequently happen arrays in data sets by scrutinizing the connotation between diagnosing and treatments. Famous Indian scientists in medical records Balakrishnan and Narayanaswamy **[34]** deliberated the qualities of choosing Persecution with Support Vector Machine to categorize the diabetics and diabetes data sets. Saraee et.al **[35]** well mined the drugs and health consequence by approving the Fuzzy Cognitive Maps in their research.

The work done by them has a way to improve the call sustenance and scheming methods in health care spheres to formulate a lot of queries, which also known as a Knowledge

Discovery Question Language and has been used to determine the hidden and applicable knowledge in medical data since then. Their work and research has surveyed a magnificent ways to mine the medical knowledge and shows the bright side to the medical experts while working on medical data mining.

III. Data processing and methods

A. Aimed hospitals

For data collection, we have aimed the seven hospitals in 'Nigeria' **[36]**, cause 'Nigeria' is the developing state of Africa, where people are struggling with the basic life needs **[37]** and they carry along the diabetes with unknown facts. These hospitals are:

1. Ajingi general hospital.
2. Gaya general hospital.
3. Sir Sanusi general hospital.
4. Muhammad Jidda general hospital.
5. Murtala Muhammad specialist hospital.
6. Abdullahi Wase specialist hospital.
7. Federal Medical center Birnin-Kudu.

B. Data attributes

This section describes all the numeric and nominal attributes, which we have used in our dataset for the classification of different algorithms and techniques. These attributes are:

Age of patients (>20 and <80), Sex (male/female), Body weight in kg, Glucose level of patients, Blood pressure, Body mass index, Excessive thirst, Diabetes (Type 1, Type 2 and Gestational), Frequent urination, Weight loss/gain, Flulike symptoms, Blurred vision, Irresistibility, Slow healing, Tingling's, Skin infection, Vagal infections, Sweating, Shivering, Visual disturbance, Weakness, Hunger, Dizziness, Nervousness, Headache, Fast heart beat, Nausea, Clammy skin, Slurred speech, Drunken behavior, Drowsiness, Confusion behavior, Convulsion, Increased urination, Leg cramp, Rapid pulse, Comma, Deep rapid breath, Breath smell, Loss of appetite, Fever, Stomach, Weight loss, Fatigue, Body smell, Drowsiness, Breath shortness, High blood pressure, Concentration, Poor appetite, Vomiting, Dry/itchy skin, Spider cobwebs, Vision shadow, Blurred vision, Empty spot vision, Red vision, Eye pain, Light flashes, Straight line vision, Vision loss, Jaws pain, Perspiration, Heat intolerance, Palpitation, Tumor, Prominent eyes, Cold intolerance, Slow body function, Weight gain, Course skin, Husky skin, Body puffiness, Insulin injections intake, Cholesterol pills intake, Number of days to visit doctor lastly, Number of times visit doctor in last six months, Class (diagnosis of diabetes mellitus) tested positive or negative.

C. Classification

Classification is a technique used to assign the class labels to a dataset **[2]**. In healthcare sectors, the classification is mostly on supervised type cause the dataset are taken on the predefined test basic and class label is known in advance. Training data is sets of records, which have multi attributes, include class that is predefined. In this scenario, the model is built up as the training dataset where the model is used to assign the class to the testing dataset. Mostly in diabetes data and other disease data, the datasets are based on supervised learning **[9]** because either the data is gathered by manually as like our research or from the artificial repositories databases but the class is defined as tested positive or tested negative.

D. Data mining platform

We have used Weka (a free and non-commercial toolkit) for experiment **[38]**. Basically weka is collection of machine learning and statistical data mining algorithms, which are based on java environment. We are able to extract the useful knowledge by training the dataset through preprocessing, clustering, classification, association and the visual interface. Weka has the ability to classify dataset with different algorithms by using filters and attributes under supervised and unsupervised learning **[39]**. In past, so many experiments performed on weka for the prediction purposes to extract the hidden knowledge. But weka has introduced "Auto-weka" recently in '2017' **[38], [40]**, which we have introduced in our model to classify the real-life Diabetes dataset by evaluation on more than 150 installed classifiers in weka. As our related studies shows that all past work has done on artificial datasets by manually changing the values but we have used the real life data values for our experiment, which is good to consider the real life predictions and forecasting.

Auto-weka is a classification sub-tool that performs the selection of combined algorithms and hyper parameter optimization over the regression and classification algorithms implementation in weka to find the better accuracy on the defined dataset. Auto-weka explores hyper parameters of settings on defined dataset and gives the recommended method of classification with the high accuracy by evaluating all the classifiers and algorithms in desired time period and helps user by giving good generalization performance algorithm by saving effort in applying of different algorithms one by one. It has been available on weka package manager since '2017' as a classification algorithm also but not much worked carried out by using this model on diabetes datasets. Auto-weka has two ways of performance; one is through GUI graphic user interface weka panel and the second is by choosing it from classifier list in Classify panel. Auto-weka performs statistically rigorous evaluation internally by 10 fold cross-validation. Auto-weka basically has few options, which normally leaves as default but for better accuracy and results two options are really important; one is 'Time limit' and second is 'Memory limit'.

E. Logistic regression modeling

Logistic regression modeling is the main part of data mining to diagnose the disease and prediction on the healthcare sections **[15]**. The main purpose of our model is to forecast the prediction of diabetes growth for clinical specialists after the analysis of classification techniques. Which is possible to predict with the binary classification and logistic regression has that ability. It is important to map the data items to its default categories and classification algorithms always have that aimed to establish such models, which work based on the existing data. It is being used to predict the tendency of the dataset. In mostly cases, the variables of logistic regression have the binary classifications, which shows that logistic algorithms are always ready to solve the two-way binary classifications. By default, logistic regression model is based on the liner regression model, which can only predict the continuous values to maintain the sensitivity in the numbers field, where the values are only 0 and 1. So in proposed model the values will be 1 incase only if the value is greater than the threshold else it will be 0. Hence the, range of output works in Logistic regression is between 0 and 1. So the logistic regression adds a sigmoid function layers by equation (1.):

1. $\sigma (x)\ 1/(1+e^{(-x)}) \in [0,1]$

The sigmoid function layers summed linearly at the first stage and than predict it by using sigmoid functions as mentioned in equation (2., 3.):

2. $Pr\,(Y=+1|X) \sim \beta.X$

3. $Pr\,(Y=-1|X)=1- Pr\,(Y=+1|X)$

After analyzing our model study of logistic regression algorithms, which consist of positive group of values and the negative group of values. Every time, the variable X will be assigned to the β coefficient values, which represent the weight and in model proposed Y is indicating the patients who has diabetes. The variations between the values X and Y occurs on the bases of weight. After the accurate setup of logistic proposed model, it is easy to predict the outcome whether it is positive or negative by giving the new data input. The output results of our research model are presented in next section with name of 'proposed results'.

IV. Proposed Reults and Discussion

This section is stipulated into two parts, classification accuracy details and forecast classification results for the diabetes dataset.

A. Classification accuracy

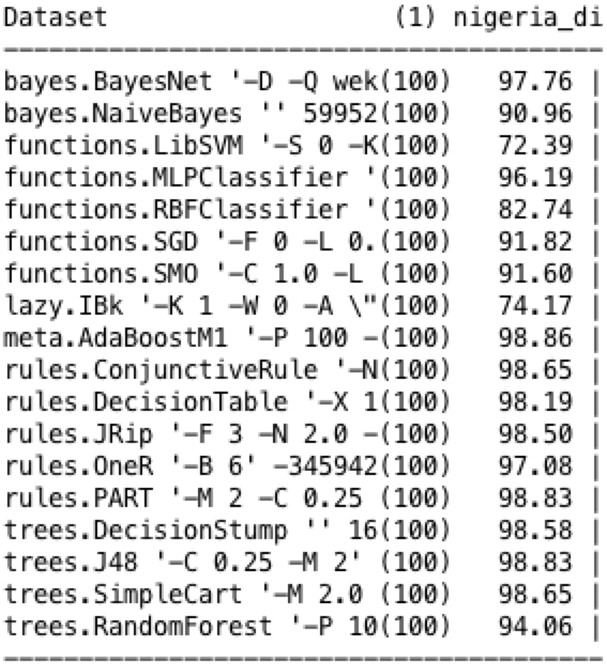
By applying our dataset of 95 attributes and 281 instances to different classification algorithms on weka in the process of auto-weka analysis with the time limit of 180 minutes to each classifier, we have gathered the anonymous results with mean errors, which are classified in Table 1.

Table 1: Shows the performance of classification algorithms performed on weka on data set of real-life diabetes patients. Acc= accuracy ratio, E.rate= error rates, K.stats= kappa statistics ratios, Pre= precision, F.msr= F measure ratio

Algorithm	Acc	E.rate	K.stats	Prec	Recall	F.msr
Bayes Net	97.5	0.03	0.94	0.97	0.97	0.97
Naïve Bayes	89.6	0.11	0.74	0.91	0.89	0.89
Lib SVM	72.2	0.27	0.13	0.80	0.72	0.63
MLP	97.1	0.07	0.93	0.97	0.97	0.97
RBF	80.7	0.33	0.49	0.81	0.80	0.79
SGD	92.8	0.07	0.83	0.92	0.92	0.92
SMO	91.1	0.08	0.78	0.91	0.91	0.91
IBK	73.3	0.26	0.37	0.73	0.73	0.73
AdaBoost	98.93	0.01	0.97	0.99	0.98	0.98
Conjective Rule	98.57	0.02	0.96	0.98	0.98	0.98
Decision Table	98.22	0.03	0.95	0.98	0.98	0.98
J Rip	97.50	0.02	0.94	0.97	0.97	0.97
One R	96.79	0.03	0.92	0.96	0.96	0.96
PART	99.28	0.07	0.98	0.99	0.99	0.99
Decision Stump	98.57	0.02	0.96	0.98	0.98	0.98
J48	99.28	0.07	0.98	0.99	0.99	0.99
Simple CART	98.57	0.02	0.96	0.98	0.98	0.98
Random Forest	93.59	0.19	0.84	0.93	0.93	0.93

B. Auto weka vs. Experimenter weka

After the analysis performed with auto weka, we have used same dataset and compared the accuracy analysis with the experimenter weka tab tool, and find the conflict ratios of results as a part of discussion to show the auto-weka as a best accuracy measure tool for the diabetes data set, the performed results of experimenter tab is shown in "Fig. 1".



```
Dataset                          (1) nigeria_di
-----------------------------------------------
bayes.BayesNet '-D -Q wek(100)     97.76 |
bayes.NaiveBayes '' 59952(100)     90.96 |
functions.LibSVM '-S 0 -K(100)     72.39 |
functions.MLPClassifier '(100)     96.19 |
functions.RBFClassifier '(100)     82.74 |
functions.SGD '-F 0 -L 0.(100)     91.82 |
functions.SMO '-C 1.0 -L (100)     91.60 |
lazy.IBk '-K 1 -W 0 -A \"(100)     74.17 |
meta.AdaBoostM1 '-P 100 -(100)     98.86 |
rules.ConjunctiveRule '-N(100)     98.65 |
rules.DecisionTable '-X 1(100)     98.19 |
rules.JRip '-F 3 -N 2.0 -(100)     98.50 |
rules.OneR '-B 6' -345942(100)     97.08 |
rules.PART '-M 2 -C 0.25 (100)     98.83 |
trees.DecisionStump '' 16(100)     98.58 |
trees.J48 '-C 0.25 -M 2' (100)     98.83 |
trees.SimpleCart '-M 2.0 (100)     98.65 |
trees.RandomForest '-P 10(100)     94.06 |
-----------------------------------------------
```

Figure 1: Shows the comparison of classification algorithms performed on Auto-weka and Experimenter weka

C. Forecast analysis

Forecasting is the approach towards the prediction and decision-making, which determines the graphical aspects of ratios including the other substances of patient's like glucose level, blood pressure and etc. According to the medical specialists **[1], [2], [11], [21], [41]–[44]**, the major substances 'attributes' involve in the prediction of diabetic patients is "body mass index, glucose level, blood pressure, body weight and insulin injection intakes", which comes on

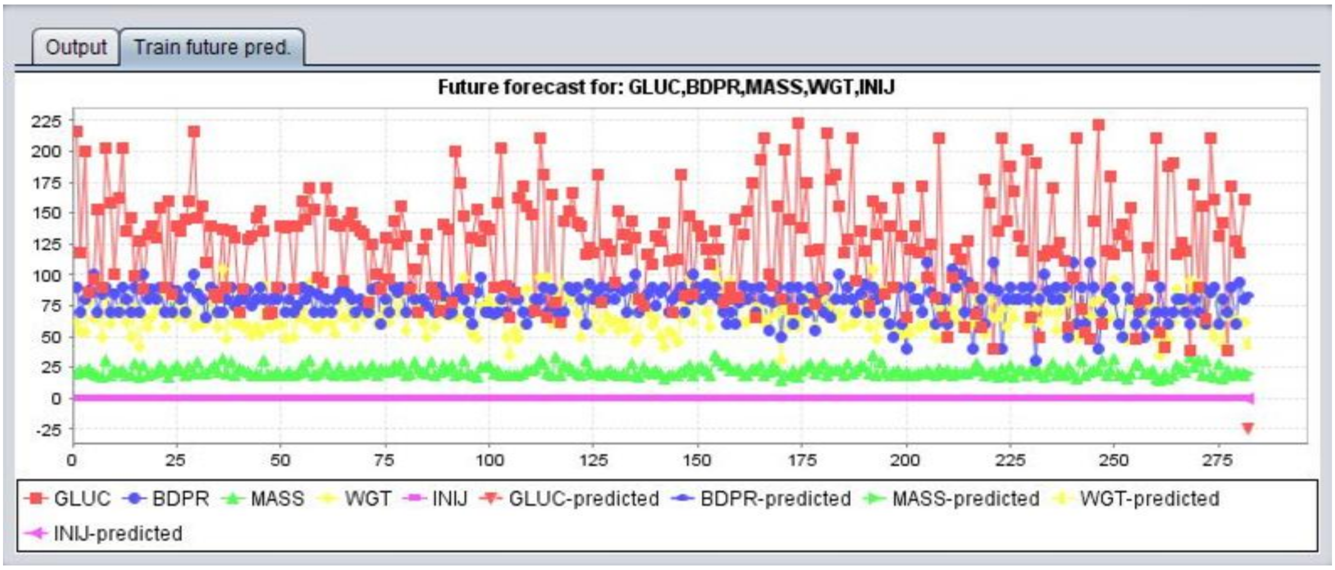


Figure 2: Forecast regression analysis of diabetes dataset attributes to predict the initial care circumstance: glucose level of patient, blood pressure, body mass index, and patient body weight and insulin injection intake

the earlier stages of patient's checkup, if the patients become predicted with these in tests of AIC, FPG, OGTT 'mentioned in above section' than the other substances 'attributes' become necessary to discover. But these 5 substances can help the medical specialist's to predict, whether the patient's has diabetes or not? If has, than the stages comes of Type 1, Type 2 and Gestational to discovers.

The graphical visualization of forecast prediction of our proposed model is shown in "Fig. 2" for the better decision-making of medical specialists.

V. Conclusion

In conclusion, this paper has aimed to establish a forecast prediction model on base of linear regression with the classification method by using weka performed on the real-life diabetes data set in order to acknowledge an education session for the medical specialists for the initial care of patients. Our results are based on the past researchers experience algorithms to perform the classification and we have pulled the relaxation accordance to free room of improvement in accuracy ratios and forecast analysis.

This methodology has only one limitation of time but it depends on the size of data only, if it gets performed on the Meta data, we have to increase the time limit of auto-weka classification methods. For the data set of 100 attributes with almost 300 instances, we have analyzed the time limit of 180 minutes is accurate in analysis.

VI. Acknowledgement

We adore prompting our appreciation to Yanshan University, Qinhuangdao, China to accompanying us in this research.

VII. References

[1] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030," *Diabetes Res. Clin. Pract.*, vol. 94, no. 3, pp. 311–321, Dec. 2011.

[2] M. N. Sohail, R. Jiadong, M. M. Uba, and M. Irshad, "A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews," *Proc. ICCD 2017*, pp. 21–26, 2017.

[3] P. Romero, Z. Obradovic, and A. K. Dunker, "Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family," *Genome Informatics*, vol. 8, pp. 110–124, 1997.

[4] Ss. Christina and N. Santiago, "Decision Support System for a Chronic Disease-Diabetes," 2018.

[5] W. Van Biesen, R. Vanholder, T. Ernandez, D. Drewniak, and V. Luyckx, "Caring for Migrants and Refugees With End-Stage Kidney Disease in Europe," *Am. J. Kidney Dis.*, vol. 71, no. 5, pp. 701–709, May 2018.

[6] S. C. Larsson, A. Wallin, N. Håkansson, O. Stackelberg, M. Bäck, and A. Wolk, "Type 1 and type 2 diabetes mellitus and incidence of seven cardiovascular diseases," *Int. J. Cardiol.*, vol. 262, pp. 66–70, Jul. 2018.

[7] M. Wei, L. W. Gibbons, J. B. Kampert, M. Z. Nichaman, and S. N. Blair, "Low Cardiorespiratory Fitness and Physical Inactivity as Predictors of Mortality in Men with Type 2 Diabetes," *Ann. Intern. Med.*, vol. 132, no. 8, p. 605, Apr. 2000.

[8] M. Godwin, M. Muirhead, J. Hyunh, B. Helt, and J. Grimmer, "Prevalence of gestational diabetes mellitus among Swampy Cree women in Moose Factory, James Bay," *CMAJ*, vol. 160, no. 9, 1999.

[9] I. Țăranu, "Data mining in healthcare: decision making and precision," *Database Syst. J.*, vol. 5, no. 4, p. 33, 2015.

[10] A. D. Association, "*Standards of Medical Care in Diabetes—2018* Abridged for Primary Care Providers," *Clin. Diabetes*, vol. 36, no. 1, pp. 14–37, Jan. 2018.

[11] M. Irshad, W. Liu, L. Wang, S. B. H. Shah, M. N. Sohail, and M. M. Uba, "Li-local: Green communication modulations for indoor localization," *Proc. 2nd Int. Conf. Futur. Networks Distrib. Syst. - ICFNDS '18*, pp. 1–6, 2018.

[12] C.-Y. Chao *et al.*, "Sleep duration is a potential risk factor for newly diagnosed type 2 diabetes mellitus," *Metabolism*, vol. 60, no. 6, pp. 799–804, Jun. 2011.

[13] X. Chen, Y. Zhang, K. Zhao, Q. Hu, and C. Xing, "Domain Supervised Deep Learning Framework for Detecting Chinese Diabetes-Related Topics," Springer, Cham, 2018, pp. 53–71.

[14] M. A. B. Monteiro *et al.*, "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, pp. 1–1, 2018.

[15] M U Muhammad, O E Asiribo, and Sohail Muhammad Noman, "Application of Logistic Regression Modeling Using Fractional Polynomials of Grouped Continuous Covariates," *Niger. Stat. Soc. Ed. Proc. 1st Int. Conf.*, vol. 1, pp. 144–147, 2017.

[16] C. M. Bennett, M. Guo, and S. C. Dharmage, "HbA $_{1c}$ as a screening tool for detection of Type 2 diabetes: a systematic review," *Diabet. Med.*, vol. 24, no. 4, pp. 333–343, Apr. 2007.

[17] A. Wallin, N. Orsini, N. G. Forouhi, and A. Wolk, "Fish consumption in relation to myocardial infarction, stroke and mortality among women and men with type 2 diabetes: A prospective cohort study," *Clin. Nutr.*, vol. 37, no. 2, pp. 590–596, Apr. 2018.

[18] M. Kanyangarara, N. Walker, and T. Boerma, "Gaps in the implementation of antenatal syphilis detection and treatment in health facilities across sub-Saharan Africa," *PLoS One*, vol. 13, no. 6, p. e0198622, Jun. 2018.

[19] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare - A Review," *Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.

[20] C. Kruse, "The New Possibilities from 'Big Data' to Overlooked Associations Between Diabetes, Biochemical Parameters, Glucose Control, and Osteoporosis," *Curr. Osteoporos. Rep.*, vol. 16, no. 3, pp. 320–324, Jun. 2018.

[21] Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention.," *Int. J. Med. Sci.*, vol. 11, no. 11, pp. 1185–200, 2014.

[22] O. Chandrakar and J. R. Saini, "Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes," in *Proceedings of the Ninth Annual ACM India Conference on - ACM COMPUTE '16*, 2016, pp. 125–128.

[23] K. Morgan, J. T. Kelly, K. L. Campbell, R. Hughes, and D. P. Reidlinger, "Dietetics workforce preparation and preparedness in Australia: A systematic mapping review to inform future dietetics education research," *Nutr. Diet.*, Jul. 2018.

[24] J. I. Wolfsdorf *et al.*, "Diabetic Ketoacidosis and Hyperglycemic Hyperosmolar State: A Consensus Statement from the International Society for Pediatric and Adolescent Diabetes," *Pediatr. Diabetes*, vol. 0, no. ja, Jun. 2018.

[25] F. Jia, L. Zuluaga-Cardona, A. Bailey, and X. Rueda, "Sustainable supply chain management in developing countries: An analysis of the literature," *J. Clean. Prod.*, vol. 189, pp. 263–278, Jul. 2018.

[26] M. Fallah and S. R. Niakan Kalhori, "Systematic Review of Data Mining Applications in Patient-Centered Mobile-Based Information Systems," *Healthc. Inform. Res.*, vol. 23, no. 4, p. 262, Oct. 2017.

[27] J. W. Maina and J. Wanjiru, "Understanding the types and causes of diabetes mellitus," *Int. J. Biol. Res.*, vol. 3, no. 1, pp. 202–207, 2018.

[28] A. M. Ahmed, "Domestic violence View project Assessment of knowledge, attitude and practice of diabetic people in Najran, Kingdom of Saudi Arabia View project," *J. Public Health (Bangkok).*, vol. 7, no. 2, pp. 56–64, 2012.

[29] S. Mirza, S. Mittal, and M. Zaman, "Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree," 2018.

[30] X. Xiong *et al.*, "ADPDF: A Hybrid Attribute Discrimination Method for Psychometric Data With Fuzziness," *IEEE Trans. Syst. Man, Cybern. Syst.*, pp. 1–14, 2018.

[31] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8102–8108, Dec. 2010.

[32] S. Chalew *et al.*, "Hemoglobin A1c, frequency of glucose testing and social disadvantage: Metrics of racial health disparity in youth with type 1 diabetes," *J. Diabetes Complications*, Aug. 2018.

[33] S. A. Mostafa, R. L. Coleman, O. F. Agbaje, A. M. Gray, R. R. Holman, and M. A. Bethel, "Modelling incremental benefits on complications rates when targeting lower HbA $_{1c}$ levels in people with Type 2 diabetes and cardiovascular disease," *Diabet. Med.*, vol. 35, no. 1, pp. 72–77, Jan. 2018.

[34] S. Balakrishnan, R. N.- Systems, M. and, and undefined 2008, "SVM ranking with backward search for feature selection in type II diabetes databases," *ieeexplore.ieee.org*.

[35] M. Saraee, G. Koundourakis, and B. Theodoulidis, "Easyminer: Data mining in medical databases," 1998.

[36] "List of Hospitals in Nigeria, State Hospital, Best Hospitals in Nigeria | VConnect^{TM}." [Online]. Available: https://www.vconnect.com/nigeria/list-of-hospitals_c289?page=1. [Accessed: 06-Sep-2018].

[37] F. Stewart, "Country experience in providing for basic needs.," *Finance Dev.*, vol. 16, no. 4, pp. 23–6, Dec. 1979.

[38] Witten, "Weka - Data Mining with Open Source Machine Learning Software in Java," *weka*, 2016. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/.

[39] S. Edition, *No Title*. .

[40] E. Sethi and K. Kumar, "A Hybrid Data Mining Approach To Evaluate Performance of Classification And Clustering Methods Implemented On Weka Platform," *© 2018 IJSRCSEIT*, vol. 5, no. 3, pp. 266–273, 2018.

[41] K. Joshi, "Survey on Different Enhanced K-Means Clustering Algorithm," *Int. J. Eng. Trends Technol.*, vol. 27, no. 4, 2015.

[42] A. Marcano-Cedeño, J. Torres, and D. Andina, "A Prediction Model to Diabetes Using Artificial Metaplasticity," Springer, Berlin, Heidelberg, 2011, pp. 418–425.

[43] P. Brambilla *et al.*, "Normal Fasting Plasma Glucose and Risk of Type 2 Diabetes," *Diabetes Care*, vol. 34, no. 6, pp. 1372–1374, Jun. 2011.

[44] A. D. American Diabetes Association, "Diagnosis and classification of diabetes mellitus.," *Diabetes Care*, vol. 37 Suppl 1, no. Supplement 1, pp. S81-90, Jan. 2014.

AUTHORS

Muhammad Noman Sohail: [Email: mn.sohail@stumail.ysu.edu.cn] is holding a B.Sc degrees (with Honors) in "Computer Networks" and M.Sc in Technology Management from "University of East London, UK" and currently pursuing PhD with project involves "Data Mining & Analysis in Health care and Fiber Sensor Technology".

Ren Jiadong: [Email: jdren@ysu.edu.cn] holds the Bachelor and Masters degree from "Harbin Institute of Technology, China". Currently he's working as a full time professor with projects involved "Data mining, Data analysis and Networks security".

Muhammad Musa Uba: [Email: musaubamuhammad@gmail.com] holds the Bachelor and Masters degree in Statistics from "Kano university of Sciences

& Technology, Nigeria "and" Ahmadu Bello University Zaria, Nigeria". Currently he's enrolled in PhD with "Information Technology projects".

M. Irshad: [Email: Ibrahim@stumail.ysu.edu.cn] holds M.Sc degree from "University of Punjab, Lahore, Pakistan". Currently he's doing PhD with project involves "Sensor Networks and data analysis".

Musavir Bilal: [Email: musavir@stumail.ysu.edu.cn] holds B.Sc Honors degree in "Computer System Engineering" from "Islamia University, Pakistan". Currently he's is doing M.Sc in "Optical Engineering" specializing in "Fiber Sensor Technology"

Usman Akbar: [Email: usman.akbar@stumail.ysu.edu.cn] is currently pursuing PhD from "Yanshan University, China" under school of Economics and Management.

Tahir Rizwan: [Email: tahirrizwan@stjtu.edu.cn] is currently pursuing PhD in Control and Engineering from "Shanghai Jiaotong University, Shanghai. China".