An Expert Diabetes Patients Prediction Support System using Modified Random Forest Classification Technique

Amanpreet Kaur, Prabhjeet kaur

Abstract-Customer acquisition and customer retention are one of the most competitive factors in most of the companies. Due to ever increasing competitions of customers in companies, the company owners are unable to maintain the customer satisfaction which leads to customer churn. The customer wishes to leave the service of the company causes churn. Most of the sectors are affected by churn problems. Telecommunication is one of the main industries that are affected by churn problem. Prediction of customers who are at risk of leaving a company is known as churn prediction and it is imperative for sustainable growth of a company. Supervised classification technique suits best for solving this problem. In this research work a technique is proposed using Correlation based Symmetric uncertainty feature selection and ensemble learning for customer churn prediction in telecommunication industry. Dataset has been collected from UCI Dataset repository and various other telecommunication websites. To predict whether a customer will be a churner or non-churner, there are a number of data mining techniques applied for churn prediction, such as artificial neural networks, decision trees, and support vector machines. The proposed customer churn prediction technique for telecommunication industry is tested using various parameters like accuracy, precision, recall, f-measure and error rate. Proposed technique is compared with support vector machine technique and Bayesian network. Experimental results demonstrate that proposed technique outperforms various algorithms in certain parameters and also gives fewer errors as compared to the previous techniques.

Keywords — Churn prediction, customer churn, churner, non-churner, customer acquisition, telecommunication industry.

I. INTRODUCTION

The major challenge faced by the various healthcare organizations like high technology hospitals and many medical centers is the delivery of standard services at cheaper costs which can be afforded by ever individual. The decisions made in scientific environment that supply optimistic outcomes are fully is dependent healthcare professional’s notion and proficient know-how reward in clinical databases. Some different types of choices are always taken which may be in fact now not a just right unsympathetic resolution and can result in catastrophic results that are intolerable. Most of sanatoriums use their scientific knowledge to analysis sufferers and producing confident results on patients. The potential hidden within the clinical databases procures us to making a terrible resolution in patient analysis. Medical resolution help integrated with laptop generated patient documents would enhance sufferer safeguard, scale back medical blunders, make stronger victim outcome and reduce damaging practice variant. This hidden understanding may utilize to analysis a sufferer who's affected by exact ailments e.g. Coronary heart ailments and in addition predict the ailment on the basis of symptoms.

Any specialist can utilize shrouded data to treat his patient. Finding of confirmation from learning extricate from clinical databases is incredible employment in front medicinal people. EBM utilizes an information mining innovation that makes it conceivable to consequently investigate colossal clinical Databases and to find designs behind them [1]. The combination of confirmation based solution rules into clinical choice emotionally supportive networks would both enhance quality and lessen expenses of care, by suggesting rules for just the most proficient medicines and drugs. Inner clinical involvement in coordination with the outer clinical aptitude must be open to the human services authorities at the fitting time and in the proper way.Diabetes is one of the most normal non-transmittable infections around the globe. It is evaluated to be the fourth or fifth purpose behind death in most high-salary nations. Diabetes is designated by IDF (International Diabetes Federation) as a standout amongst the most difficult medical problems of 21st century. Diabetes is known malady since antiquated circumstances. Diabetes is an ailment that is a metabolic issue in which individual has high hints of glucose in the blood caused by deficient generation of glucose content in the body on the grounds that the body cells don't react the way they need to really react to insulin [16]. In the event that the follow level of glucose increments in the blood then it will be indicated by the different manifestations, for example, overwhelming thirst, visit pee, unexplained weight reduction and so on. [16]. There are many data mining techniques to predict the diabetes analysis.

II. BACKGROUND

ShravanKumar et al. [2] outlines a specialist framework that predicts the coronary illness and diabetes sickness. The creator utilizes decreased number of traits and after that utilizations information mining method in which he connected C4.5 order calculation so that there is more exactness and less run time. The creator likewise applies choice tree calculation for the expectation of coronary illness and diabetes and prognosticates that whether infection is available or not. As indicated by the creator the current technique takes 0.05 sec while the proposed strategy took around 0.025 sec and the precision is additionally expanded from 84.35% to 85.96%.

A. Iyer, et al. [3] searched answers for analyze the malady by breaking down the examples found in the information through arrangement examination by utilizing Decision Tree and Naïve Bayes calculations. The programmed analysis of diabetes is a vital certifiable medicinal issue. Identification of diabetes in its beginning times is the key for treatment. This paper indicate how Decision Trees and Naïve Bayes are utilized to display real determination of diabetes for neighborhood and efficient treatment, alongside introducing related work in the field. The performance of the techniques was investigated for the diabetes diagnosis problem. Experimental results demonstrate the adequacy of the proposed model.

Veena Vijayan et al. [4] proposed a choice emotionally supportive network that utilizations AdaBoost calculation with Decision Stump as base classifier for characterization. Moreover Support Vector Machine, Naive Bayes and Decision Tree are additionally executed as base classifiers for AdaBoost calculation for exactness confirmation. The exactness got for AdaBoost calculation with choice stump as base classifier is 80.72% which is more prominent contrasted with that of Support Vector Machine, Naive Bayes and Decision Tree. Different mechanized data frameworks were sketched out using various classifiers for foreseeing and diagnosing diabetes. Choosing true blue classifiers plainly extends the precision and capability of the framework.

Amit kumar et al. [5] used various techniques like Artificial Neural Network, K-fold cross validation and classification, Vector support machine, K-nearest neighbor method, Data Mining Algorithm for diagnosis of diabetes and attempted to make an ensemble model by combining two techniques: Bayesian classification and Multilayer Perceptron for the accuracy, sensitivity and specificity measures of diagnosis of diabetes-mellitus. In this experiment, author used various individuals and hybrid classification models for classification of diabetes data. The analysis of models is done in two steps: first model is trained and tested. Various data mining techniques like C4.5, random forest (RF), Bayes Net and Multi Layer Perceptron (MLP) are trained using randomly training data set and after that the testing of the trained models is done using randomly tested data set. Partitions of data plays very important role in accuracy of models. Accuracy is varying from partition to partition.

Thirumal P. C. et al.[6] applied different information mining methods which are imperative to forecast of diabetes mellitus and extricate concealed examples from the PIMA Indian diabetes dataset accessible at UCI Machine Learning Repository. The utilization of information mining methods in ailment expectation is to decrease the test and increment the precision of rate of location. A standout amongst the most widely recognized illnesses among youthful grown-up is Diabetes mellitus. In this paper, a few information mining calculations, for example, Naïve Bayes, Decision trees, k Nearest neighbor and SVM have been examined and tried with pima Indian diabetes dataset.

Srideivanai Nagarajan et al. [7] aimed to outline and execute a specialist clinical framework to analyze the sort of diabetes and the levels of hazard among diabetic patients utilizing the information mining methods bunching and grouping. The examination configuration made utilization of essential and optional information and the information were gathered utilizing information accumulation devices and systems, for example, polls, coordinate meeting and review of existing restorative records from 650 diabetic patients. The arrangement calculations - NaiveBayes, Random Tree, Simple Cart and Simple Logistic were utilized on the bunched information to group the information into gentle, direct and serious composes coming about into a specialist clinical framework.

Tahani Daghistani et al. [8] applied three information mining calculations, to be specific Self-Organizing Map (SOM), C4.5 and Random Forest, on grown-up populace information from Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia to foresee diabetic patients utilizing 18 hazard factors. Arbitrary Forest accomplished the best execution contrasted with other information mining classifiers. The informational collections are gathered from Ministry of National Guard Health Affairs (NGHA) databases from the most noteworthy three populated areas in Saudi Arabia, where the databases have all patients visit data, for example, research center and solutions, and so forth. The outcomes demonstrate that the built information mining model could help social insurance suppliers to settle on better clinical choices in distinguishing diabetic patients. Moreover, the model could be additionally produced for quiet security.

Sajida Perveen et al. [9] followed the adaboost and bagging troupe methods utilizing J48 (c4.5) choice tree as a base student alongside independent information mining procedure J48 to group patients with diabetes mellitus utilizing diabetes hazard factors. This arrangement is done crosswise over three diverse ordinal grown-ups bunches in Canadian Primary Care Sentinel Surveillance organize. This investigation built sensibly great models with higher execution to characterize diabetic patients, crosswise over three age bunches in the Canadian populace, utilizing sacking adaboost and additionally J48 choice tree. The dataset utilized as a part of this investigation is acquired from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database. Assessment of results shows that adaboost outfit strategy beats than packing and in addition independent J48 choice tree.

B. Senthil Kumar et al. [10] analyzed and compared diverse information mining and machine learning strategies utilized as a part of diabetes. The assignment of malady conclusion and anticipation is a piece of characterization and forecast. The current and prevalent information mining methods utilized as a part of clinical information incorporates Bayesian, Random timberland calculations, Artificial Neural system, SVM and Decision Tree and so on.

Panigrahi Srikanth et al. [11] evaluated Classification Algorithms for the Classification of some Diabetes Disease Patient Datasets. This paper anticipated Diabetes Disease in view of Data Mining Techniques of Classification Algorithms. Order Algorithm and devices may lessen substantial work on Doctors. Characterization Algorithm analyzes the Decision Tree Algorithm, Byes Algorithm and Rule based Algorithm. These calculations assess Error Rates and recognize patients in light of development capacity to quantify the precise outcomes.

Ekta et al. [12] considered holistic approach to analyze and classify the diabetes dataset for data preprocessing. In order to do so, diabetes.arff dataset is used for data preprocessing and prediction of diabetes. From this research work one can easily analyzed that WEKA tool is quite useful for analyzing the given dataset. Finally, it is analyzed that persons who are suffered from diabetes have age more than 40 and mass more than 35. On the other hand, the persons who are not infected from diabetes have age less than 30 and mass less than 35.

P. Suresh et al. [13] proposed a model to take care of the issues in existing framework in applying information mining systems to be specific grouping and orders which are connected to analyze the sort of diabetes and its seriousness level for each patient from the information gathered. This paper attempted to analyze diabetes in light of the 650 patient's information with which creator investigated and distinguished seriousness of the diabetes. As a major aspect of method Simple k-implies calculation is utilized for grouping the whole dataset into 3 bunches i.e., bunch 0 - for gestational diabetes, bunch 1 for type-1 diabetes (adolescent diabetes), bunch 2 for type-2 diabetes. This grouped dataset was given as contribution to the order show which additionally orders every patient's hazard levels of diabetes as gentle, direct and extreme. Further, execution investigation of various calculations has been done on this information to analyze diabetes.

Ioannis Kavakiotis et al. [14] aimed to lead an orderly audit of the utilizations of machine learning, information mining strategies and instruments in the field of diabetes explore concerning a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the main classification giving off an impression of being the most well known. Support vector machines (SVM) emerge as the best and generally utilized calculation. Concerning the kind of information, clinical datasets were chiefly utilized. The title applications in the chose articles venture the convenience of separating profitable information prompting new theories focusing on more profound understanding and further examination in DM.

S.Selvakumar et al. [15] predicted the persons whether diabetic or not. In this paper classification techniques such as Binary Logistic Regression, Multilayer Perceptron and K-Nearest Neighbor are classified for diabetes data and classification accuracy were compared for classifying data. Using data mining methods to aid people to predict diabetes has gain major popularity. This work focused the implementation of Binary Logistic Regression, Multilayer Perceptron and k-Nearest Neighbor for the diabetes data. From the analysis, it is examined that the formation of classifications will be different for classification methods. From the histogram, it is seen that the Binary Logistic Regression accuracy is 0.69, Multilayer Perceptron accuracy is 0.71 and KNN gives the accuracy of 0.80.k- Nearest Neighbor is higher than the accuracy of Binary Logistic Regression and Multilayer Perceptron.

Messan Komi et al. [16] explored the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN. The diabetes prediction system is developed using five data mining classification modeling techniques. These models are trained and validated against a test dataset. All five models are able to extract patterns in response to the predictable states. The most effective model to predict patient with diabetes appear to be ANN followed by ELM and GMM. Although not the most effective model, the Logistic regression result is easier to read and interpret, what is more, the training over Logistic regression is very efficient. Although the ANN do outperform other data mining methods, the relationship between attributes and the [mal result is more difficult to understand.

Gauri D et al. [17] implemented machine learning calculation in Hadoop MapReduce condition for Pima Indian diabetes informational index to discover missing esteems in it and to find designs from it. The calculations can attribute missing esteems and to perceive designs from the informational collection. Prescient investigation is a strategy that coordinates different information mining procedures, machine learning calculations and insights that utilization present and past informational collections to find learning from it and by utilizing it foresee future events.

III.PROPOSED TECHNIQUE

The abundant amount of knowledge hidden in clinical databases that can be efficiently utilize to diagnosis of patient's diseases. This hidden information present in clinical databases can be used as guidelines after long

process of lab research, specialist verification and positive result with patient diseases. In the existing paper, diabetics patients have been analyzed using single decision tree. This will lead to less accurately classified analysis.
In the proposed approach, number of decision trees are combined for the analysis process in order to make a forest and will improve the accuracy.

The proposed approach comprises of a hybrid clustering with classification approach i.e. Weighted Hierarchical clustering and Balance Random forest on Diabetes patient's datasetand analyse the predictions.

The methodology for proposed technique is as follows:
Collection of data: Dataset were mainly collected from UCI repository and from various hospitals of Hemodialysis disease. Patient's data were collected which contains 8 attributes.

Preprocessing and Filtering: In preprocessing step, it chooses a property for choosing a subset of trait with the goal that it can give great anticipated capacity. It additionally contains the transformation of information composes like numeric to ostensible or the other way around. It handles all the missing esteems and expel them. On the off chance that a characteristic contains over 5% missing esteems, at that point the records ought not be erased and it is encouraged to put the qualities where the information is missing utilizing some reasonable techniques and aides in highlight determination and class name.

Classification: Classification is a system for machine learning by which it is utilized to foresee the gathering participation of various information occasions. It will play out the assignment by which it will sum up the outstanding structure in order to apply it on new information. Here Random backwoods classifier has been utilized for quality estimation of dataset will be consider based on level of effectively ordered occasions. For approval stage we utilize 10-overlay cross approval technique. Irregular backwoods classifier helps in recognizing the attributes of patient with Diabetes illnesses.

Hierarchical clustering
Let X = {x1, x2, x3, ..., xn} be the set of data points.
1) Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
2) Find the least distance pair of clusters in the current clustering, say pair (r), (s), according to d[(r),(s)] = d[(k), (r,s)] = ($ulp((Var)^{-1}\ \sum_{i=1}^{n}(x_i - y_i)^2)$)). where, $Var\ =\ \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2)}{n-1})\bar{x}\ is\ the\ mean\ of\ attributes$ where the minimum is over all pairs of clusters in the current clustering.
3) Increment the sequence number: m = m +1.Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)].
4) Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (r,s) and old cluster(k) is defined in this way:
d[(k), (r,s)] = ($ulp((Var)^{-1}\ \sum_{i=1}^{n}(x_i - y_i)^2)$)). where, $Var\ =\ \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2)}{n-1})\bar{x}\ is\ the\ mean\ of\ attributes$
5) If all the data points are in one cluster then stop, else repeat from step 2).

Balanced Random Forest
1. For each iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement, from the majority class.
2. Induce a classification tree from the data to maximum size, without pruning. The tree is induced with the C4.5 Tree algorithm, with the following modification: At each node, instead of searching through all variables for the optimal split, only search through a set of m- try randomly selected variables.
3. Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final prediction.

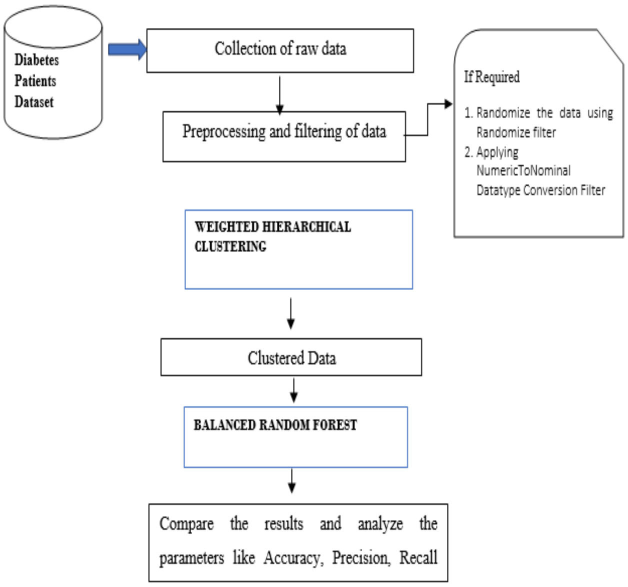
Diabetes Patients Dataset
Collection of raw data
Preprocessing and filtering of data
If Required
1. Randomize the data using Randomize filter
2. Applying NumericToNominal Datatype Conversion Filter
WEIGHTED HIERARCHICAL CLUSTERING
Clustered Data
BALANCED RANDOM FOREST
Compare the results and analyze the parameters like Accuracy, Precision, Recall

Figure 1: Flow chart of the proposed technique

IV. EXPERIMENTAL RESULTS

The simulation has been done in Java Net Beans. NetBeans is an open-source project dedicated to providing rock solid software development products (the NetBeans IDE and the NetBeans Platform) that address the needs of developers, users and the businesses who rely on NetBeans as a basis for their products; particularly, to enable them to develop these products quickly, efficiently and easily by leveraging the strengths of the Java platform and other relevant industry standards.

Dataset Used

The dataset utilized as a part of this exploration work is gathered from National Institute of Diabetes and Digestive and Kidney Diseases and depends on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of machine learning databases. The Pima Indian diabetes database, gave by Vincent Sigillito, is an accumulation of medicinal demonstrative reports of 768 illustrations. Before 1694, they alluded to themselves as OTAMA. The name PIMA is thought to have gotten from correspondence issues amongst Europeans and individuals from the Otama clan.

Number of Instances: 768

Table 1: Diabetes dataset attributes

1	Number of times pregnant	Real	3.8	3.4
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Real	120.9	32.0
3	Diastolic blood pressure (mm Hg)	Real	69.1	19.4
4	Triceps skin fold thickness (mm)	Real	20.5	16.0
5	2-Hour serum insulin (mu U/ml)	Real	79.8	115.2
6	Body mass index (weight in kg/(height in m)^2)	Real	32.0	7.9
7	Diabetes pedigree function	Real	0.5	0.3
8	Age (years)	real	33.2	11.8

Performance metrics

i. **Precision and recall**
Precision and recall are the two metrics that are widely for evaluating performance in text mining, and in text analysis field like information retrieval. These parameters are used for measuring exactness and completeness respectively.

$$\text{Precision}=\frac{\text{True Positive}}{\text{True Positive+False Positive}} \text{Eq. (1)}$$

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad \text{Eq. (2)}$$

ii. **F-measure**
F-Measure is the harmonic mean of precision and recall. The value calculated using F-measure is a balance between precision and recall.

$$\text{F measure} = \frac{2*\text{recall}*\text{precision}}{\text{precision+recall}} \quad \text{Eq. (3)}$$

iii. **Accuracy**
Accuracy is the common measure for classification performance. Accuracy can be measured as correctly classified **instances** to the total number of **instances**, while error rate uses incorrectly classified instances instead of correctly classified instances.

$$\text{Accuracy}=\frac{\text{True Positive+True Negative}}{\text{True Positive+False Positive+True Negative+False Negative}} \text{Eq. (4)}$$

Table 2: Representing the Accuracy of proposed method with respect to previous methods

Algorithm	Accuracy
KNN	64.58
Decision Tree	65.1
Random Forest	60.29
Proposed Modified HC with Random Forest	98.96

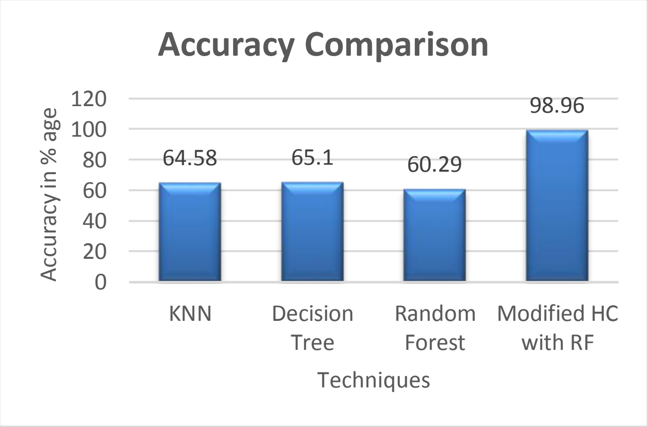

Figure 2: Showing the accuracy comparison of existing with the proposed classification algorithm

The figure above shows the accuracy comparison of the existing algorithms including KNN, Decision Tree, Random Forest and proposed Modified HC with Random Forest algorithm. The graph clearly shows that the proposed algorithm performs better as its accuracy is 98.96% i.e. it has more efficiently classifies the patients

Table 3: Accuracy and Inaccuracy comparison of Existing and Proposed Classification on Diabetes Dataset

Algorithm	Accuracy	Inaccuracy
KNN	64.58	35.42
Decision Tree	65.1	34.9
Random Forest	60.29	39.71
Proposed Modified HC with Random Forest	98.96	1.04

The table above shows the accuracy and inaccuracy comparison between the algorithms. The algorithm with more accuracy and less inaccuracy is the better among all the above. The table clearly shows that the proposed Modified HC (Hierarchical Clustering) with random Forest is better than the existing classification algorithms.

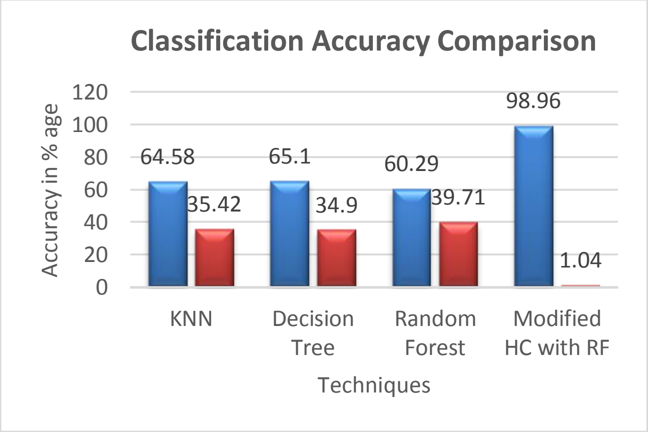

Figure 3: Showing the accuracy and Inaccuracy comparison of existing with the proposed classification algorithm

The figure above shows the accuracy and Inaccuracy comparison of the existing algorithms including KNN, Decision Tree, Random Forest and proposed Modified HC with Random Forest algorithm. The graph clearly shows that the proposed algorithm performs better as its accuracy is 98.96% and Inaccuracy is 1.04 which is very less as compared to other algorithms.

Table 4: Class details parameters comparison of Existing and Proposed Classification on Diabetes Dataset

Parameters	KNN	Decision Tree	Random Forest	Modified HC with Random Forest
Precision	0.636	0.424	0.536	0.979
Recall	0.646	0.651	0.603	0.99
F-Measure	0.64	0.513	0.545	0.984

The table above shows the class parameters comparison between the algorithms. The algorithm with more precision, recall and Fmeasure is the better among all the above. The table clearly shows that the proposed Modified HC (Hierarchical Clustering) with random Forest is better than the existing classification algorithms.

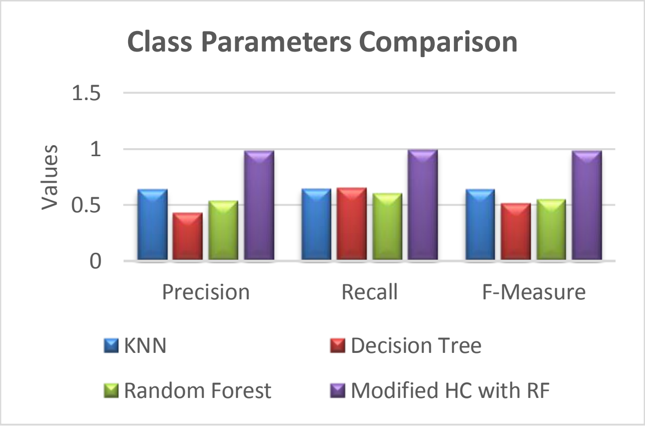

Figure 4: Showing the class parameters comparison of existing with the proposed classification algorithm

The figure above shows the class parameters comparison of the existing algorithms including KNN, Decision Tree, Random Forest and proposed Modified HC with Random Forest algorithm. The class parameters include precision, recall and Fmeasure. The graph clearly shows that the proposed algorithm performs better as its precision, recall and Fmeasure all are more than the existing base classifiers.

Table 5: Error rate comparison of Existing and Proposed Classification on Diabetes Dataset

Error rate	KNN	Decision tree	Random Forest	Modified HC with Random Forest
Mean	0.3885	0.4544	0.3961	0.0105

absolute error				
Root mean square error	0.5247	0.4766	0.5689	0.0763

The table above shows the error rate comparison between the algorithms. The algorithm with less error rate i.e. mean absolute error and Root mean square error is the better among all the above. The table clearly shows that the proposed Modified HC (Hierarchical Clustering) with random Forest is better than the existing classification algorithms.

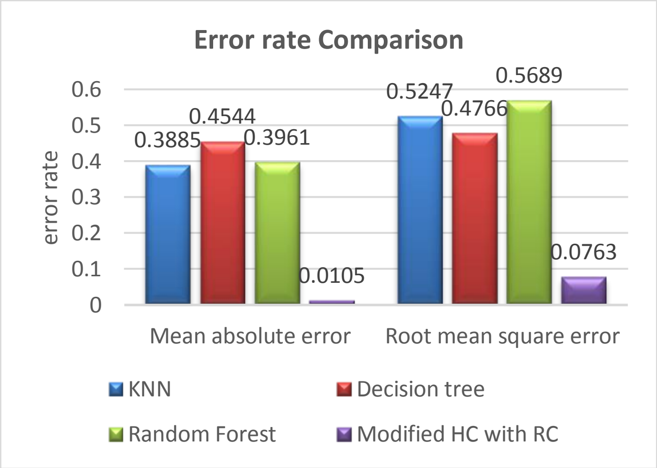

Figure 5: Showing the error rate comparison of existing with the proposed classification algorithm

The figure above shows the error rate comparison of the existing algorithms including KNN, Decision Tree, Random Forest and proposed Modified HC with Random Forest algorithm. The class parameters include precision, recall and Fmeasure. The graph clearly shows that the proposed algorithm performs better as its error rate is less than the existing base classifiers.

V. CONCLUSION

In this research work, number of decision trees are combined for the analysis process. The proposed technique applies Weighted Hierarchical clustering and Balance Random forest on Diabetes patient's dataset. The automatic diagnosis of diabetes is an important real-world medical problem. Detection of diabetes in its early stages is the key for treatment. Dataset were mainly collected from UCI repository and from various hospitals of Haemodialysis disease. Patient's data were collected which contains 8 attributes. The simulation has been done in Java Net Beans. Results are compared using Diabetes dataset on proposed algorithm in contrast with the existing algorithms including KNN, Decision Tree and Random Forest. Results shows that the proposed algorithm classify the diabetes patients more efficiently by showing highest accuracy. Proposed technique gives accuracy of 98.96, 0.979 precision, 0.99 recall, 0.984 Fmeasure. Proposed technique gives very less error rate of 0. 0105.

In future, the use data mining technique on any other medical condition. Here, we perform clustering on classification. In future, feature selection may be combined with classification so that clustering time is reduced as feature selection optimizes the features and selects the best features.

REFERENCES

[1] Candice MacDougall, Jennifer Percival and Carolyn McGregor (2009), "Integrating Health Information Technology into Clinical Guidelines", IEEE Annual International Conference, 2009, pp. 4646-4649.

[2] ShravanKumar Uppin and M A Anusuya (2014), "Expert System Design to Predict Heart and Diabetes Diseases", International Journal of Scientific Engineering and Technology, Volume No.3, Issue No.8, 2014, pp : 1054-1059.

[3] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process, Vol.5, No.1, 2015, pp. 1-14.

[4] Veena Vijayan V., Anjali C., "Prediction and Diagnosis of Diabetes Mellitus -A Machine Learning Approach", IEEE Recent Advances in Intelligent Computational Systems, 2015, pp. 122-127.

[5] Amit kumar Dewangan, Pragati Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques", International Journal of Engineering and Applied Sciences, Volume-2, Issue-5, 2015, pp. 145-148.

[6] Thirumal P. C. and Nagarajan N, "Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus - A Case Study", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 1, 2015, pp. 8-13.

[7] Srideivanai Nagarajan and R. M. Chandrasekaran, "Design and Implementation of Expert Clinical System for Diagnosing Diabetes using Data Mining Techniques", Indian Journal of Science and Technology, Vol 8(8), 2015, pp. 771–776.

[8] Tahani Daghistani, Riyad Alshammari, "Diagnosis of Diabetes by Applying Data Mining Classification Techniques", International Journal of

Advanced Computer Science and Applications, Vol. 7, No. 7, 2016, pp. 329-332.

[9] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes”, Science Direct Symposium on Data Mining Applications, 2016, pp. 115-121.

[10] B. Senthil Kumar, Dr. R. Gunavathi, “A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 12, 2016, pp. 463-467.

[11] Panigrahi Srikanth, Dharmaiah Deverapal, “A Critical Study of Classification Algorithms Using Diabetes Diagnosis”, IEEE 6th International Advanced Computing Conference, 2016, pp. 245-249.

[12] Ekta, Sanjeev Dhawan, “Classification of Data Mining and Analysis for Predicting Diabetes Subtypes using WEKA”, International Journal of Scientific & Engineering Research, Volume 7, Issue 12, 2016, pp. 100-103.

[13] P. Suresh Kumar and V. Umatejaswi, “Diagnosing Diabetes using Data Mining Techniques”, International Journal of Scientific and Research Publications, Volume 7, Issue 6, 2017, pp. 705-709.

[14] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research”, Elsevier Computational and Structural Biotechnology Journal 15, 2017, pp. 104–116.

[15] S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar, “Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques”, International Journal of Statistics and Systems, Volume 12, Number 2, 2017, pp. 183-188.

[16] Messan Komi, J un Li, Y ongxin Zhai, Xianguo Zhang, “Application of Data Mining Methods in Diabetes Prediction”, IEEE 2nd International Conference on Image, Vision and Computing, 2017, pp. 1006-1010.

[17] Gauri D. Kalyankar, Shivananda R. Poojara, Nagaraj V. Dharwadkar, “Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, IEEE International conference on I-SMAC, pp. 619-624.