Big Data Analytics in Health Care by Data Mining and Classification Techniques

T.Velumani*
Assistant Professor, Department of Computer Science,
Rathinam College of Arts and Science (Autonomous), Coimbatore, Tamilnadu, India
Corresponding Author
Email Id: velumani46@gmail.com

ABSTRACT

Big data is the compilation of intricate and enormous quantity of data that arrives as of unrelated cause for instance online operation information, social media, sensor data, etc. Such assortment of huge data develop into tough to evaluate by conventional processing relevance's. By the development and upcoming latent in the field of healthcare business, it is essential to analyze a huge quantity of noisy data to get significant information. In healthcare system, the intend of this work is to estimate the health record of diabetes patients by a combination of innovative hierarchical decision attention network, association rules (AR) and multiclass outlier classification with MapReduce framework. The association rule apriori algorithm in a MapReduce framework considers health data to create regulations. This is employed to discover the association among disease and their signs. This examination is made by means of UCI machine learning datasets of diabetes containing 50 attributes. The results of the proposed algorithm are offered by parameters for instance precision, accuracy, recall, and F-score.

Keywords: *Big data; data mining; Apriori algorithm (AA); Map Reduce; parallel computing.*

INTRODUCTION

Big data analytics (BDA) is an emerging topic among scholars and practitioners and is termed as a holistic scheme to supervise, practice and analyze the 5 V data- associated magnitudes (*i.e.*, velocity, volume, variety, veracity and value) [6]. BDA has comprised in various applications including healthcare units, business and industrial sectors [9]. The high volume advanced surge of data that is being produced at ever-higher velocities and assortments in healthcare include complexity. The consequences are redundant boost in health costs and moment for patients and healthcare facility suppliers [7]. These consequences include technological obstacles with hierarchical, social, economic, and policy boundaries [11]. A mass of information crosswise all phases, from infection conclusion to analgesic concern, is additional suggestion of the chances and test to useful information administration, investigation, forecast and optimization procedure as division of information administration in medical situations [12].

Diabetes mellitus (DM) referred as diabetes is a long-standing metabolic chaos in which the blood glucose (BG) rank differ and is rooted by any inadequate insulin manufacture in the body (sort 1 diabetes, T1D) or by the body's powerlessness to exploit its formed insulin (sort 2 diabetes, T2D) [15]. Based on the results of the Diabetes difficulties and manage check, stretched glycemic manage is essential to lessen the threat of long-standing microvascular and macrovascular difficulties, together with nephropathy, retinopathy, neuropathy and cardiovascular illness, in people with sort 1 DM [14]. Therefore, healthcare associations are looking for effective IT artifacts that authorize them to consolidate authoritative resources to

convey a high quality patient experience [8]. So far a few works are proposed to tackle BDA in human services issues and it includes the introduction of a scalable MapReduce-rooted logistic relapse to route colossal quantify of sensor information [10].

Diverse machine learning (ML) algorithms are utilized to recognize a most excellent forecast algorithm like Random Forest (RF), Support Vector Machine (SVM), K-means Nearest Neighbor (k-NN), Classification and Regression Trees (CART) and Linear Discriminant Analysis (LDA) algorithms [13]. On the other hand a predictive investigation algorithm in the Hadoop/MapReduce condition is proposed to forecast the DM difficulties and the sort of healing to be embraced [16]. Consequently a system that qualities the predictive examination of algorithm in Hadoop/Map Reduce condition to predict and classify the sort of Diabetic Mellitus, Type-1 diabetes doesn't produce the insulin in our body [17].

To rectify these issues a major information technique is association rule mining which is indented to decide fascinating associations among things and to build up a set of AR used as in [18]. This is a technique for discovering dependencies that attest that the assignment of identifying association reliance's in human services databases [19]. Mining AR centre on attaining associations amid data [20].The work is prepared as pursue. Section 2 momentarily depicts several of the existing approaches in detail. Section 3 briefly examines the proposed MapReduce framework (MRF) and Section 4 analyzes outcome. After all, Section 5 presents conclusions and future extent.

RELATED WORK FOR THE RESEARCH

Some of the recent research works which are proposed to rectify the issues with BDA in healthcare are discussed below:

Kumar *et al.* [1] Described about large information which is the collection of complex and gigantic measure of information. Such gathering of huge information develops into rigid to split down utilizing customary processing applications. Doctors approved the insulin to the Diabetic patients and the choice depends on the patient's past documentation and measure the sugar stage at the customary interims. The point of this work is to break down the medicinal database of diabetes patients utilizing information mining algorithms. To play out the examination, the stages such as Hadoop and mapreduce are utilized notwithstanding the information mining algorithms, i.e., decision tree (DT) and Naive Bayes. This investigation is finished utilizing uci ML datasets of Diabetes including four features for the preparation stage.

Namrata Bhattacharya *et al*. [2] Described healthcare production, where it is essential to break down a lot of boisterous information to get significant information. Information mining techniques were practical to expel incoherent information and mine important examples. Association rule mining (ARM) is a standard rooted technique which reveals how things are linked with all other. AA is a comprehensively utilized algorithm for digging successive item sets for ARM. Be that as it may, the presentation of the AA corrupts with the enormous quantity of information. Here, an execution of the AA was given in Hadoop MapReduce structure. Medical information was considered to construct rules which might be utilized to discover the association among sickness and their side effects. These standards can be utilized for information finding to give rules to the healthcare business.

Chen *et al*. [3] explained the lifelong and systematic injure suffered by diabetes patients, which is critical to design effective tactics for the finding and healing of diabetes. In view of

comprehensive assessment, this work classifies Diabetes 1.0 and Diabetes 2.0 strategies, which display shortages as far as networking and intelligence. In this manner, a low cost, and the shrewd diabetes finding result with customized healing was created. Right now, the 5G-Smart Diabetes scheme was proposed, which unites the best in class technologies for instance wearable 2.0, ML, and enormous information to create broad detecting and examination for patients experiencing diabetes. At that point the information sharing method and customized information investigation representation were introduced for 5G-Smart Diabetes.

Bai *et al*. [4] Discussed about exponentially developing information and BDA as a rising pattern. Information mining procedure assumes a brisk job in the appliance of Big Data in the healthcare division. DM algorithms provide a presentation to examine, discover and forecast the existence of malady and aid doctors in DM by premature discovery and correct administration. The principle objective of information mining methods in healthcare scheme is to plan a mechanized device which analyzes the health information and underwear the patients and doctors regarding the force of the malady and the sort of cure to be preeminent practiced dependent on the side effects, persistent documentation and healing record. This work underlines on diabetes medicinal information where classification and clustering algorithms are executed and the competence of the equivalent is analyzed.

Moreira *et al.* [5] discussed that the advancement of keen decision emotionally supportive networks (DSSs) that look to reenact human social characteristic as a chief dispute for computational intelligence (CI). Artificial neural network (ANN) approaches can take care of complex decision-production issues. The application of this transformative CI method to dissect a lot of information is a significant methodology to get concern of a few issues in healthcare administration. This work proposed the modelling, presentation assessment, and evaluation investigation of an ANN technique called the Radial Basis function network (RBF Network) to recognize potential cases of gestational diabetes that can prompt various dangers for both the pregnant ladies and the baby. This research gives a broad decision-production representation proficient of getting better the concern gave to ladies who are at a danger of creating gestational diabetes, which is the mainly frequent metabolic issue in development with a frequency of 3–18%. Consequently, this effort contributed to the decrease of maternal and fetal death and horribleness rates.

Rawal, Bhavna, and Ruchi Agarwal [23] executed C4.5 to support up the classification accuracy and move reverse planning to construct a classification representation. Shankar *et al*. [24] gave "DataSpeak" an approach for mining, classification, and conglomeration in large information. This approach overcome the impediment of kNN and furnishes quick right to use of information with well-organized information mining and classification in a negligible time. Bechini *et al*. [25] introduced a scattered association rule-based classification plan moulded presenting to the MapReduce programming model. The method extracts distribution AR (CARs) utilizing an appropriately improved, scattered adaptation of the notable FP-Growth algorithm.

Fernández *et al*. [26] broke down the interrelation between the number of marks of the fuzzy factors and the scarcity of the information because of the information inspecting in MapReduce. Specifically, it also considered the apportioning of the underlying instance set develops the level of granularity etc. Elkano *et al*. [27] gave a distributed Fuzzy Rule-Based Classification Systems (FRBCSs) as Big Data classification issues called CHI-BD. This strategy depends on the notable Chi et al. algorithm and utilizes Big Data structure, i.e., Apache Hadoop. Game *et al*. [28] proposed a major information classification model

(coronary illness) in social insurance, which comprises definite stages. The means are as per the following: (1) Map-reduce structure (2) SVM (3) streamlined DT classifier (DT).
Banchhor, Chitrakant, and N. Srinivasu [29] built up a classifier, called Cuckoo–Gray wolf based Correlative Naive Bayes classifier (CG-CNB), by transforming CNB classifier with a recently evolved optimization algorithm, and Cuckoo–Gray Wolf based Optimization (CGWO). Sivaparthipan, C. B., N. Karthikeyan, and S. Karthik. [30] Proposed a model of a arithmetical evaluation, healthcare data scheme for Diabetes Analysis utilizing large information. Saru, S., and S. Subashree [31] discovered the information mining techniques and strategies to locate the fitting approaches and methods for efficient classification of Diabetes dataset and in extracting important examples.

Chen, Peihua, and Chuandi Pan [32] introduced two general boosting algorithms; Adaboost.M1 and LogitBoost, to create a apparatus representation for diabetes analysis placed on the above mentioned clinical test data. Sohail *et al.* [33] established a prediction model for DM prevalence for the accuracy measurement ratios by applying 150 ML classifiers and for the improvement in clustering with positive and negative instances. Rghioui *et al.* [34] discussed classification methods are employed for forecast on the dataset of patient's data, to analyze general diabetic presentation and predict some relative's disease.
Younus *et al.* [35] proposed ML approaches such as DT, RF for developing classification system-rooted forecast representation to measure sort DM chronic diseases. Moreover, an algorithm was given rooted on RF to see the difficult regions of sort diabetes patients. Pani *et al.* [36] explored two classifier training methods logistic regression (LR) and SVM. Sampath *et al.* [37] used the analytical investigation algorithm in the Hadoop/MapReduce background to guess the diabetes difficulties and the categorization of healing to be approved.

Though these several existing methods are identified further enhanced classification accuracy is not achieved. Hence the proposed hierarchical classification algorithm is comprised of the following contributions:

1) Initially a Map reduce framework is implemented since the dataset utilized is too big. This Mapreduce comprises of a mapper function and a reducer function, thus reducing the quantity of data whereas preserving significant data for instance anatomical appropriate data.
2) An innovative hierarchical decision attention network algorithm where the DT is processed due to its simplicity that formulates a education representation and anticipate the worth of a target attribute by learning the conclusion of the inferred decision rules. Additionally the hierarchical attention network utilized with DT and is applied to the word and sentence-level, facilitating it to focus further and not as much of significant content while creating the document illustration.
3) Then AA that uses AR is completed that analyzes frequent itemset data and by means of the criterion of support and confidence to discover the recurring associations amid data called associations.
4) Finally, this work does an outlier based multiclass classification which classifies based upon the prediction made with the support and the confidence score in AA.

PROPOSED HIERARCHICAL ALGORITHM

DM is one of the developing lethal ill everywhere all through the world. Due to the evolution of the electronic machines in health care large quantity of data is obtained leading to broad patient population, noisy data, heterogeneous data, *etc*. This leads healthcare BDA as an evolving research topic. In recent healthcare scheme, huge dealing out of shapeless physical condition files, to access these files for examination and make it dynamic behind

investigation, will build additional complications. Due to this, health industry faces many difficulties to analyze this kind of data and there is a necessitate to extend the data analytics. The intend of mining information is to obtain out the data from the dataset and build a lucid image for its supplementary exploit. Diabetes is a widespread disease which is on elevated climb and most of the individuals suffering from this disease in the entire earth. The patient containing the diabetic dilemma has a count of side effects for instance stroked, nerve damage, heart disease, kidney damage, *etc*. Therefore, it is necessary to notice this dilemma early on which secures the patient from former fitness problems. The overall aim of this work is to forecast the occurrence of diabetes datasets to forecast the optimal result for the patients rooted on the results accurately.

Doctors decide the insulin based on the patient's earlier documentation and determine the sugar stage at the usual periods. For this examination, the platforms for instance Hadoop and MapReduce are employed in accumulation to the innovative data mining (DM) algorithms, i.e., DT attention network, AR and outlier based multiclass classification technique will be proposed. Though existing methods work with this Hadoop and mapreduce platforms they worked with only less number of attributes. So this method will provide better results by considering many attributes of patients. By this proposed hierarchical technique the diabetic patient database will be analyzed effectively since after mapreduce platform a classification technique called multiclass outlier technique. Consequently, based upon the classification results obtained insulin will be given in appropriate amount considering patient record and diabetic results.

Hadoop and Map Reduce

Hadoop technology is utilized for storing, accessing and analyzing large sources of big data. It employs the MRF which works with parallel processing (PP) of data. MapReduce is a PP programming which is applied for PP above the collection of nodes called clusters. It improves the idea of PP on thousands of servers during MapReduce. A mapper and reducer are two phases to be done in Hadoop. The yield of map phase is given as input to reduce. As the MRF is rooted on the key value set-up a mapper gets the input and split the mined data into key–value pairs as the output. The input to the mapper phase is a diabetes .csv file. Here, initially certain numbers of field’s or attributes are extracted from every record. The gone value in the dataset is also identified. In mapper stage, all documentation processed in succession and separately in parallel and creates pairs of key–value and is equated as

$$Map(key1, value1) -> list(key2, value2) \quad (1)$$

Alternatively, reduce phase allows the output of mapper phase as its input and obtains the key–value pairs in text correspondingly. Hence, the assessments of diabetes to the count of patients fit in to a scrupulous surveillance date at the specified time and specific conditions are monitored. The final output of the reduce stage is the united structure of all intermediate values which are gathered subsequent to processing and frequently in the mixture of key–value pairs. Subsequent to every phase of processing, the output acquires arranged in an organized way and is equated as

$$Reduce(key2, list(value2)) -> list(key3, value3) \quad (2)$$

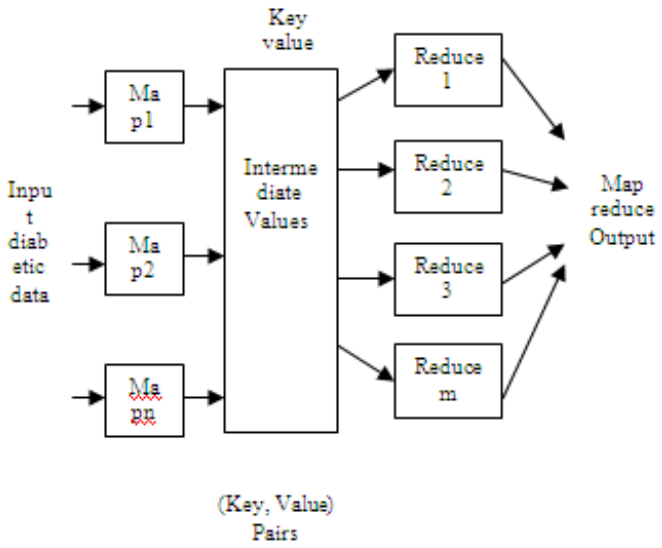


Fig. 1. Map Reduce Framework

The detailed processing within the MRF is pictorially portrayed in fig. 1.From MapReduce the big data dataset is reduced into a smaller dataset and hence preceded for further processing. The output of MapReduce algorithm is comprised of different attributes.

Hierarchical Decision Attention Network Algorithm

Subsequently the result from the MapReduce algorithm which is a minimized big data table is fed into a hierarchical decision attention network.

Decision Tree

A DT algorithm is a supervised education algorithm appropriate for classifying categorical data by their attributes. It exploits a DT approach to devise a learning model which can estimate the value of a target attribute by learning the results of the decision rules incidental from the training data. Here the classification rules are signified by paths from root to leaves. In DM, the MapReduce data are represented as in (3)

$$(u,V)=(u_1,u_{2,}u_3,.....,u_n,V) \quad (3)$$

V is the target attribute that is to be classified with aid of the values of input variables or attributes like x1, x2, x3, *etc*., represented by vector x.

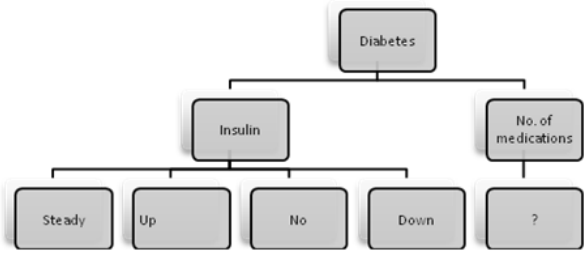


Fig. 2. Decision Tree Classifier

The process of DT classification is outlined in fig. 2. Right now presence of insulin is detected by different qualities where insulin and a number of medications are considered important and are represented. Possibly the utmost advantage of DTs is their simplicity.

Hierarchical Attention Network

The Hierarchical Attention Network [22] consist of an encoder which typically processes input information with a recurrent layer such as LSTM, and a decoder which maps the encoded contribution to the ideal yield, typical with a second recurrent layer. This also permits the decoder to centre on parts of the encoded input while creating the interpretation. The encoder comprises of an installing that converts categorical tokens into numeric vectors followed by two LSTM operations. The decoder comprises a consideration mechanism different from the encoder. The consideration mechanism permits the decoder to ensure of specific pieces of the encoder yield.

Alternatively, the DT results are used by this hierarchical consideration network, in this manner shaping a hierarchical decision consideration network. The info table is divided for preparing information and testing information. At that point the information, content information is converted to numeric sequences and spaces are embedded between the characters. Sequence information, for example content normally has different sequence lengths. To prepare a model utilizing variable length sequences, cushion smaller than normal batches of info information to have a similar length. To guarantee that the cushioning esteems don't impact the misfortune calculations, create a veil which records which sequence components are genuine, and which are simply cushioning. The yield sequences have different lengths, so they require cushioning. The corresponding cushioning, the cover contains zeros where the corresponding time steps are cushioning esteems. At that point the yield is taken care of as contributing to the AA.

Apriori Algorithm

The yield of the DT consideration network algorithm is given to AA which utilizes AR. The AR break down continuous itemset information and utilizing the criteria of help and confidence to locate the recurring connections amid information called associations. Hence AR is implication articulations of the structure as in (4)

$$\{u_1 \rightarrow other\ attributes\ from(u_2,...u_n)\} \tag{4}$$

The support gives a gauge of the possibility of happening of an occasion, *i.e.* $P\{u_1 \rightarrow other\ attributes\ from(u_2,...u_n)\}$. The help of the itemset is detailed by condition (5). If the help of an itemset is more noteworthy than or equivalent to the minimum_support, at that point that itemset is added to the arrangement of incessant itemsets.

$$\sup port = \frac{No\ of\ transactions}{No\ of\ transactions\ in\ database} \tag{5}$$

The confidence of the rule $P\{u_1 \rightarrow other\ attributes\ from(u_2,...u_n)\}$ specifies the possibility of both antecedent and resultant showing up in a similar transaction. Confidence is planned by condition (6). If the confidence of the inference prepared by a standard is more noteworthy than or equivalent to the minimum_confidence, at that point that standard is added to the arrangement of AR.

$$Confidence\{u_1 \rightarrow other\ attributes\ from(u_2,...u_n)\} = \frac{\sup port(u_1 \cup (u_2,...u_n))}{\sup port(u_1)} \quad (6)$$

Among numerous applications of ARM, one of the significant applications is in sickness investigation, which includes the mapping of ailment to their treatment. The standards acquired from AA are then taken care of as contribution to the following stage, for example exception based multiclass classification technique.

Outlier Based Multiclass Classification

The yield of AA is then taken care of to exception based multiclass classification technique [21]. The exception based multiclass classification is classified into two sorts such as multi-class and in class. This work is performed multiclass classification by which the AA yield is classified depending on the prediction made with the help and confidence scores. The dataset is part in preparing and testing tests. Beginning preparing is done to detect the diabetic patients. Hence testing is finished by which the corresponding classification of diabetic patient is made. The algorithm ventures for the proposed hierarchical classification algorithm is point by point beneath.

Algorithm 1 : Proposed hierarchical classification algorithm
Input : Diabetic big data set
Output **:** Classified result

1) Begin
2) Load the diabetic dataset
3) Perform Mapreduce algorithm
4) Map the dataset
5) Reduce the dataset with less attributes and records
6) The DT algorithm starts
7) Produce decision
8) Run attention network
9) Run the AA
10) Generate set of rules
11) Run outlier based multiclass classification
12) Classify the result
13) By classification result
14) If
15) Diabetes = yes
16) Then
17) Give insulin = {up, down, steady}
18) Else if
19) Diabetes = no
20) Then
21) Give insulin = {no}
22) End

SIMULATION RESULTS AND DISCUSSIONS

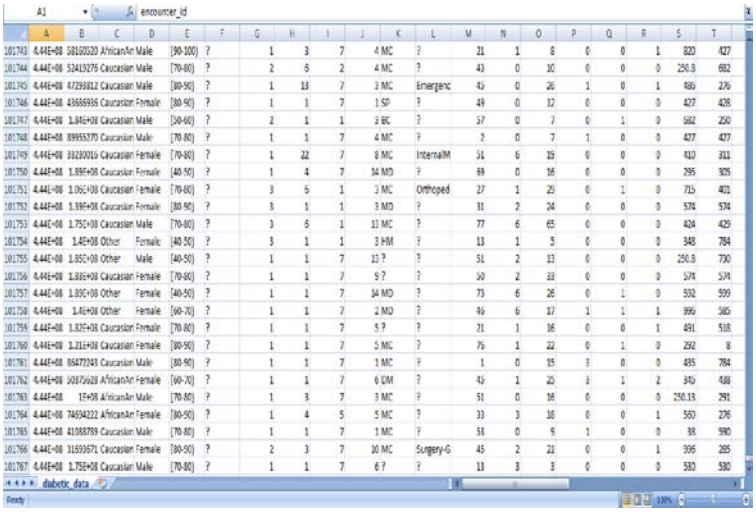
The dataset was engaged from https://archive.ics.uci.edu/ml/datasets/diabetes UCI ML. This diabetic dataset comprises of 50 attributes and is tabulated in table 1.

Table 1: Attributes in Diabetic Dataset

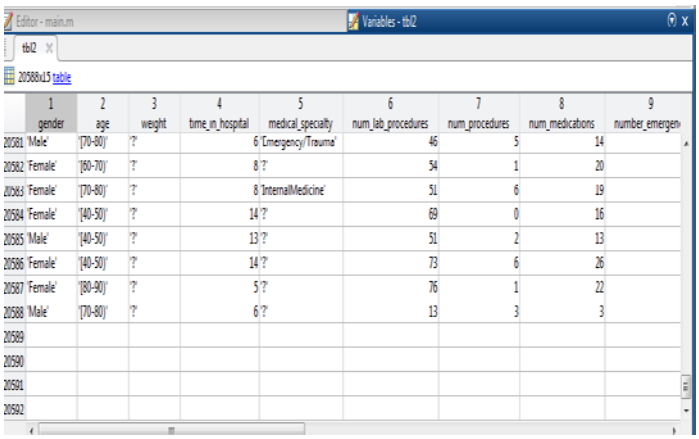
encounter_id	patient_nbr	Race
Gender	age	Weight
admission type id	discharge disposition id	admission_source_id
time_in_hospital	payer_code	medical_specialty
num_lab_procedures	num_procedures	num_medications
number_outpatient	number_emergency	number_inpatient
diag_1	diag 2	diag 3
number_diagnoses	max_glu_serum	A1Cresult
Metformin	repaglinide	Nateglinide
Chlorpropamide	glimepiride	Acetohexamide
Glipizide	glyburide	Tolbutamide
Pioglitazone	rosiglitazone	Acarbose
Migitol	troglitazone	Tolazamide
Examide	citoglipton	Insulin
glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone
metformin-rosiglitazone	metformin-pioglitazone	Change
diabetesMed	readmitted	

Thus, the dataset comprises of 50 attributes with 101767 records. The experimental study is approved out on the base of the implementation of proposed hierarchical decision attention network, association rule and multiclass outlier classification algorithm by MapReduce technique and results are calculated. The output is executed in Matlab 2019 a with 64 bit operating system environment.

Fig. 3 clearly views the information about the count of records in the original diabetic dataset and the count of records in the mapreduce results. By this it is clearly shown that the big data diabetic dataset is reduced to a great extend. That is along with the records the attributes are also reduced. Hence, after mapreduce processing there are only 15 attributes in the dataset such as the age, gender, medical_specialty, weight, time_in_hospital, num_lab_procedures, number_emergency, num_procedures, num_medications, number_diagnoses, max_glu_serum, metforim, insulin, dibetesMed and readmitted. Further, this mapreduce dataset is utilized for additional processing and is fed as input to a hierarchical decision attention network.



(a) Original diabetic dataset with 101767 records



	1 gender	2 age	3 weight	4 time_in_hospital	5 medical_specialty	6 num_lab_procedures	7 num_procedures	8 num_medications	9 number_emergen
20581	'Male'	'[70-80)'	'?'	6	'Emergency/Trauma'	46	5	14	
20582	'Female'	'[60-70)'	'?'	8	'?'	54	1	20	
20583	'Female'	'[70-80)'	'?'	8	'InternalMedicine'	51	6	19	
20584	'Female'	'[40-50)'	'?'	14	'?'	69	0	16	
20585	'Male'	'[40-50)'	'?'	13	'?'	51	2	13	
20586	'Female'	'[40-50)'	'?'	14	'?'	73	6	26	
20587	'Female'	'[80-90)'	'?'	5	'?'	76	1	22	
20588	'Male'	'[70-80)'	'?'	6	'?'	13	3	3	
20589									
20590									
20591									
20592									

(b) Dataset records obtained after Mapreduce

Fig 3. Result illustration for mapreduce framework

```
Parallel mapreduce execution on the parallel pool:
*********************************
*      MAPREDUCE PROGRESS       *
*********************************
Map    0% Reduce    0%
Map   50% Reduce    0%
Map  100% Reduce    0%
Map  100% Reduce  100%
```

Fig 4. Mapreduce Parallel execution

The output of the DT algorithm obtained is depicted in fig. 5.

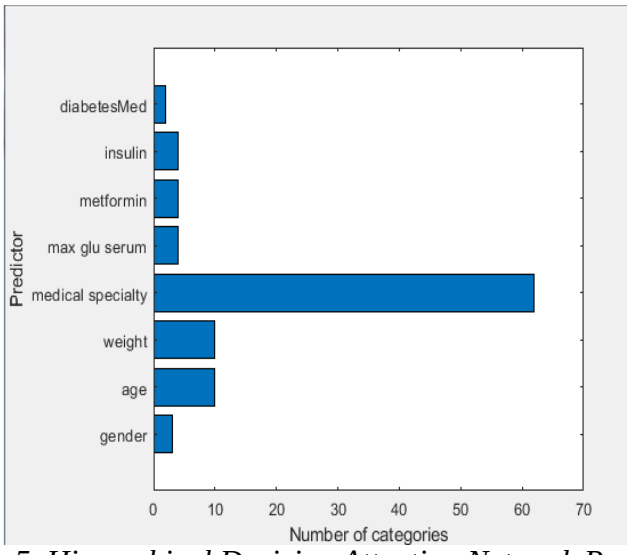


Fig. 5. Hierarchical Decision Attention Network Results

An association rule AA is applied to disease analysis and hence a patient who experience from diabetes should be given insulin. ARM is employed to determine implications such as {insulin} → {diabetes}. These rules are explained pictorially in fig. 7 and also formulated in equation (7) as.

$$\{insulin, age\} \Rightarrow \{diabetesMed\} \quad (7)$$

Also the support value is identified as 0.40 and that of the confidence value is identified as 0.67.

Fig. 6. Apriori Algorithm rule prediction

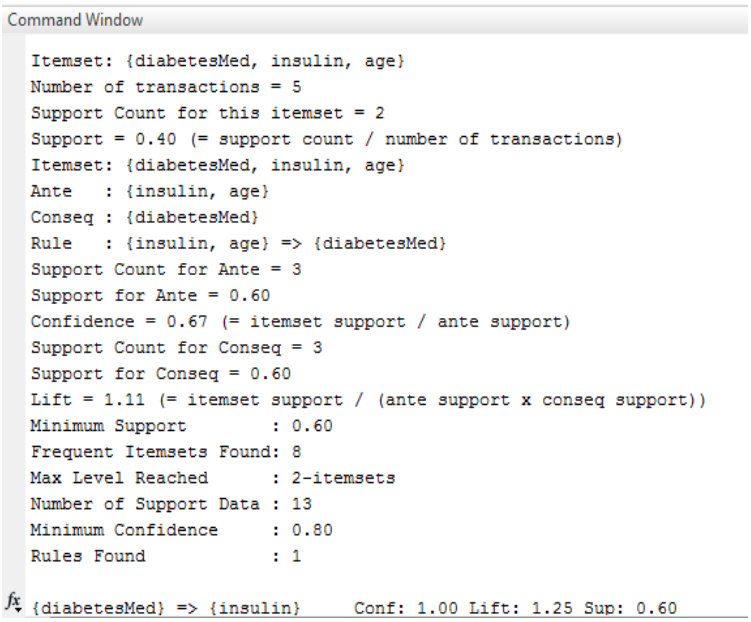


Fig. 7. Apriori Support and Confidence Results

The result of classifying is predicted into two class issues such as correctly and incorrectly. The confusion matrix has conditions like accuracy, Probability of detection (sensitivity), true negative rate (specificity), precision, F1score. Higher value of 'True Positive' detection is enviable for diabetic classification. Accuracy is formulated as the percentage of the count of diabetic patients classified correctly versus total the diabetic patient as in (8). Probability of detection or TPR or sensitivity or recall measures the proportion of actual diabetic patients who is correctly identified and is formulated in (9). True Negative rate (TNR) or specificity measures the proportion of non-diabetic patients who are properly recognized and is equated in (10). Precision is defined as the proportion of accurately forecasted diabetic patients to the forecasted size of the diabetic patients and is formulated as in (11). Consequently the classification accuracy is estimated by the subsequent formulas

$$Accuracy = (TP + TN) \Big/ (TP + FP + FN + TN) \quad (8)$$

$$\text{Pr}obability\ of\ \text{det}ection = TP \Big/ (TP + FN) \quad (9)$$

$$True\ negative\ rate = (TN) \Big/ (FP + TN) \quad (10)$$

$$\Pr ecision = TP \Big/ (TP + FP) \quad (11)$$

where True Positive (TP) are the count of diabetic patients classified as an diabetic patient. True Negative (TN) is the count of normal classified as normal. False Positive (FP) is the number of normal classified as diabetic patient and False Negative (FN) is the count of diabetic patients classified as normal.

Table 2 : Comparison of the Proposed Methodology

Parameters	Proposed Hierarchical algorithm	DT + apriori + outlier multiclass	DT	RBF Network
Precision	1	0.99	0.89	0.824
Probability of detection	1	0.7	0.65	0.871
F-score	1	0.82	0.73	0.847

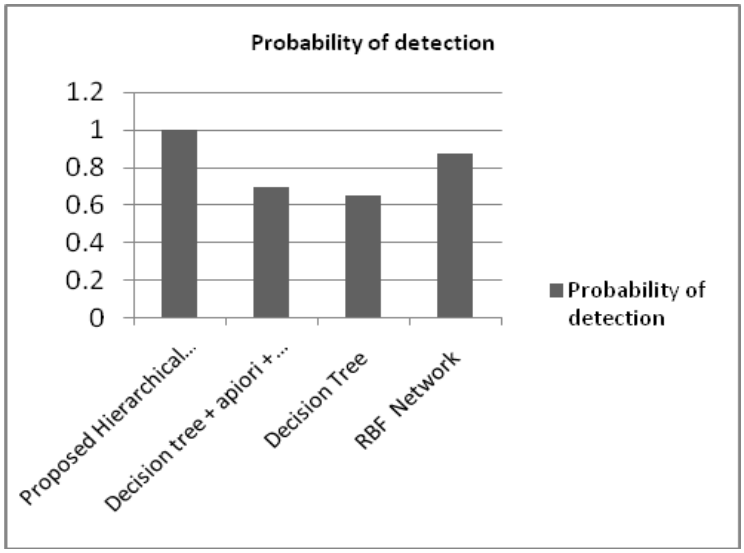


Fig 8. Recall Comparison Graph

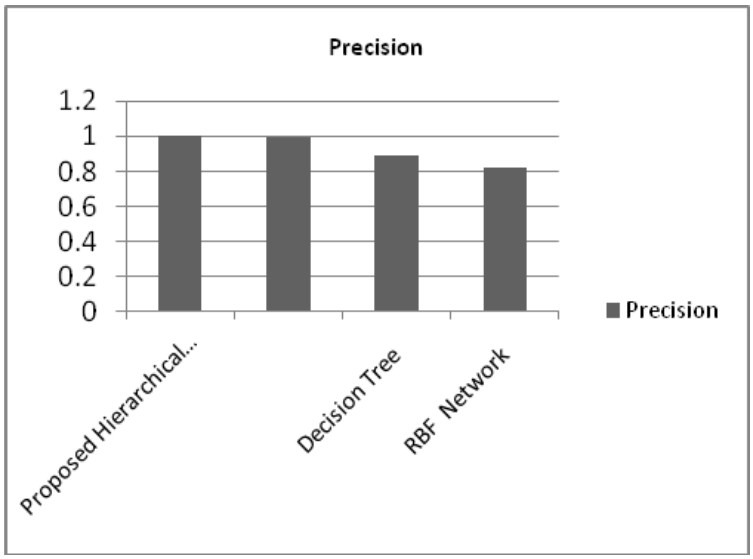


Fig. 9. Precision Comparison Graph

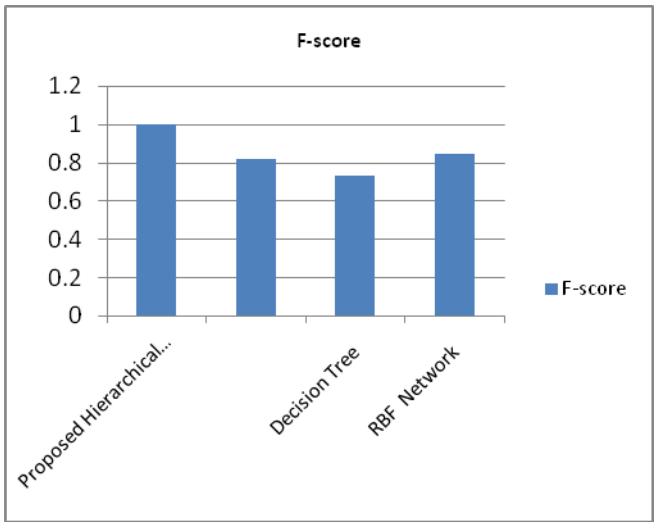


Fig. 10. F-score Comparison Graph

RBF Network is utilized to recognize the probable cases of gestational diabetes that can direct to various threats for both the pregnant women and the fetus. This technique guesses the unidentified values using neurons that employ radial base activation functions. This approach used the k-means clustering algorithm to grant 255 the basis functions and study by performing a linear regression on the output layer of the network to forecast possible gestational diabetes risk cases. Hence it does not show improved results when compared to the proposed hierarchical algorithm.

In the given graph in fig. 9 with the diabetic dataset, it demonstrates that the precision value of the proposed hierarchical algorithm, i.e., 1.is more than as compared to the DT algorithm i.e. 0.99 and that of the RBF Network value i.e. 0.824. Fig. 8 proves the recall value comparison where the proposed hierarchical method has higher value of compared when compared to a DT algorithm which is 0.7 and that of the RBF Network value which is 0.871. Fig. 10 shows the F-score evaluation of the proposed hierarchical method with existing methods for instance DT and RBF Network.

CONCLUSION AND FUTURE WORK

The aim of this work is to forecast more accurately the occurrence of diabetes datasets to discover the optimal solution for the patient. To perform this analysis, the MapReduce platforms are used in addition to the proposed hierarchical algorithm such as hierarchical decision attention network, AA and outlier based multiclass classification. By this the diabetes patient is classified and there insulin levels are determined. From the comparison graph depicted it is obvious that the proposed hierarchical algorithm shows improved performance. The performance indicators of confusion matrices used are precision, recall and F-score, *i.e.* 1 are executed on the dataset. In the future, this algorithm will be allowed to cloud computing structure for improved access and perform in real time.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1) Kumar, Sunil, and Maninder Singh. "Diabetes Data Analysis Using Mapreduce with Hadoop." *In Engineering Vibration, Communication and Information Processing*, Springer, Singapore, 2019, pp. 161-176.

2) Bhattacharya, Namrata, Sudip Mondal, and Sunirmal Khatua. "A MapReduce-Based Association Rule Mining Using Hadoop Cluster—An Application of Disease Analysis." *In Innovations in Computer Science and Engineering*, Springer, Singapore, 2019, pp. 533-541.
3) Chen, Min, Jun Yang, Jiehan Zhou, Yixue Hao, Jing Zhang, and Chan-Hyun Youn. "5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds." *IEEE Communications Magazine, vol.* 56, no. 4, pp. 16-23, 2018.
4) Bai, BG Mamatha, B. M. Nalini, and Jharna Majumdar. "Analysis and Detection of Diabetes Using Data Mining Techniques—A Big Data Application in Health Care." *In Emerging Research in Computing, Information, Communication and Applications*, Springer, Singapore, 2019, pp. 443-455.
5) Moreira, Mário WL, Joel JPC Rodrigues, Neeraj Kumar, Jalal Al-Muhtadi, and Valeriy Korotaev. "Evolutionary radial basis function network for gestational diabetes data analytics." *Journal of computational science*, vol. 27, pp. 410-417, 2018.
6) Wamba, Samuel Fosso, Angappa Gunasekaran, Shahriar Akter, Steven Ji-fan Ren, Rameshwar Dubey, and Stephen J. Childe. "Big data analytics and firm performance: Effects of dynamic capabilities." *Journal of Business Research*, vol. 70, pp. 356-365, 2017.
7) Belle, Ashwin, Raghuram Thiagarajan, S. M. Soroushmehr, Fatemeh Navidi, Daniel A. Beard, and Kayvan Najarian. "Big data analytics in healthcare." *BioMed research international* ,2015
8) Wang, Yichuan, LeeAnn Kung, and Terry Anthony Byrd. "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations." *Technological Forecasting and Social Change*, vol. 126, pp. 3-13, 2018.
9) Loebbecke, Claudia, and Arnold Picot. "Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda." *The Journal of Strategic Information Systems*, vol. 24, no. 3, pp. 149-157, 2015.
10) Manogaran, Gunasekaran, Daphne Lopez, Chandu Thota, Kaja M. Abbas, Saumyadipta Pyne, and Revathi Sundarasekar. "Big data analytics in healthcare Internet of Things." *In Innovative healthcare systems for the 21st century*, Springer, Cham, 2017, pp. 263-284.
11) Kankanhalli, Atreyi, Jungpil Hahn, Sharon Tan, and Gordon Gao. "Big data and analytics in healthcare: introduction to the special section." *Information Systems Frontiers*, vol. 18, no. 2, pp. 233-235, 2016.
12) De Silva, Daswin, Frada Burstein, Herbert F. Jelinek, and Andrew Stranieri. "Addressing the Complexities of Big Data Analytics in Healthcare: The Diabetes Screening Case." *Australasian Journal of Information Systems*, vol.19, 2015.
13) Kumar, P. Suresh, and S. Pranavi. "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics." In 2017 *International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, IEEE, 2017, pp. 508-513.
14) Prahalad, P., M. Tanenbaum, K. Hood, and D. M. Maahs. "Diabetes technology: improving care, improving patient‐reported outcomes and preventing complications in young people with Type 1 diabetes." *Diabetic Medicine*, vol. 35, no. 4, pp. 419-429, 2018.
15) Alfian, Ganjar, Muhammad Syafrudin, Muhammad Ijaz, M. Syaekhoni, Norma Fitriyani, and Jongtae Rhee. "A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing." *Sensors*, vol. 18, no. 7, pp. 2183, 2018.

16) Sampath, P., S. Tamilselvi, NM Saravana Kumar, S. Lavanya, and T. Eswari. "Diabetic data analysis in healthcare using Hadoop architecture over big data." *International Journal of Biomedical Engineering and Technology*, *vol.* 23, no. 2-4, pp. 137-147, 2017.
17) Prasad, S. Thanga, S. Sangavi, A. Deepa, F. Sairabanu, and R. Ragasudha. "Diabetic data analysis in big data with predictive method." *In 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, IEEE, 2017, pp. 1-4.
18) Abdel-Basset, Mohamed, Mai Mohamed, Florentin Smarandache, and Victor Chang. "Neutrosophic association rule mining algorithm for big data analysis." *Symmetry*, vol. 10, no. 4, pp. 106, 2018.
19) Shakhovska, Nataliya, Roman Kaminskyy, Eugen Zasoba, and Mykola Tsiutsiura. "Association rules mining in big data." *International Journal of Computing*, vol. 17, no. 1, pp. 25-32, 2018.
20) Chen, Yunliang, Fangyuan Li, and Junqing Fan. "Mining association rules in big data with NGEP." *Cluster Computing*, vol. 18, no. 2, pp. 577-585, 2015.
21) Ndirangu, Dalton, Waweru Mwangi, and Lawrence Nderu. "An Ensemble Model for Multiclass Classification and Outlier Detection Method in Data Mining." 2019.
22) Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical attention networks for document classification." *In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480-1489.
23) Rawal, Bhavna, and Ruchi Agarwal. "Improving accuracy of classification based on c4. 5 decision tree algorithm using big data analytics." *In Computational Intelligence in Data Mining*, Springer, Singapore, 2019, pp. 203-211.
24) Shankar, Venkatesh Gauri, Bali Devi, and Sumit Srivastava. "DataSpeak: Data extraction, aggregation, and classification using big data novel algorithm." *In Computing, communication and signal processing*, Springer, Singapore, 2019, pp. 143-155.
25) Bechini, Alessio, Francesco Marcelloni, and Armando Segatori. "A MapReduce solution for associative classification of big data." *Information Sciences*, vol. 332, pp. 33-55, 2016.
26) Fernández, Alberto, Sara del Río, Abdullah Bawakid, and Francisco Herrera. "Fuzzy rule based classification systems for big data with MapReduce: granularity analysis." *Advances in Data Analysis and Classification*, vol. 11, no. 4, pp. 711-730, 2017.
27) Elkano, Mikel, Mikel Galar, Jose Sanz, and Humberto Bustince. "CHI-BD: a fuzzy rule-based classification system for big data classification problems." *Fuzzy Sets and Systems*, vol. 348, pp. 75-101, 2018.
28) Game, Pravin S., Vinod Vaze, and M. Emmanuel. "Optimized Decision tree rules using divergence based grey wolf optimization for big data classification in health care." *Evolutionary Intelligence*, pp. 1-17, 2019.
29) Banchhor, Chitrakant, and N. Srinivasu. "Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification." *Data & Knowledge Engineering*, pp. 101788, 2019.
30) Sivaparthipan, C. B., N. Karthikeyan, and S. Karthik. "Designing statistical assessment healthcare information system for diabetics analysis using big data." *Multimedia Tools and Applications*, pp. 1-14, 2018.
31) Saru, S., and S. Subashree. "Analysis and Prediction of Diabetes Using Machine Learning." *International Journal of Emerging Technology and Innovative Engineering*, vol. 5, no. 4, 2019.
32) Chen, Peihua, and Chuandi Pan. "Diabetes classification model based on boosting algorithms." *BMC bioinformatics*, vol. 19, no. 1, pp. 109, 2018.

33) Sohail, Noman, Ren Jiadong, M. Uba, Muhammad Irshad, and Ayesha Khan. "Classification and cost benefit Analysis of Diabetes mellitus Dominance." *Int. J. Comput. Sci. Netw. Secur*, vol. 18, pp. 29-35, 2018.
34) Rghioui, Amine, Jaime Lloret, and Abedlmajid Oumnad. "Big Data Classification and Internet of Things in Healthcare." *International Journal of E-Health and Medical Communications (IJEHMC,)*, vol. 11, no. 2, pp. 20-37, 2020.
35) Younus, Muhammad, Md Tahsir Ahmed Munna, Mirza Mohtashim Alam, Shaikh Muhammad Allayear, and Sheikh Joly Ferdous Ara. "Prediction Model for Prevalence of Type-2 Diabetes Mellitus Complications Using Machine Learning Approach." *In Data Management and Analysis*, Springer, Cham, 2020, pp. 103-116.
36) Pani, Luina, Somnath Karmakar, Chinmaya Misra, and Satya Ranjan Dash. "Multilevel Classification Framework of fMRI Data: A Big Data Approach." *In Big Data Analytics for Intelligent Healthcare Management*, Academic Press, 2019, pp. 151-174.
37) Sampath, P., S. Tamilselvi, NM Saravana Kumar, S. Lavanya, and T. Eswari. "Diabetic data analysis in healthcare using Hadoop architecture over big data." *International Journal of Biomedical Engineering and Technology*, vol. 23, no. 2-4, pp. 137-147, 2017.