International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-018-00916-z

ORIGINAL ARTICLE

Random forest for big data classification in the internet of things using optimal features

S. K. Lakshmanaprabu[1] · K. Shankar[2] · M. Ilayaraja[2] · Abdul Wahid Nasir[3] · V. Vijayakumar[4] · Naveen Chilamkurti[5]

Received: 27 June 2018 / Accepted: 26 December 2018

Abstract

The internet of things (IoT) is an internet among things through advanced communication without human's operation. The effective use of data classification in IoT to find new and hidden truth can enhance the medical field. In this paper, the big data analytics on IoT based healthcare system is developed using the Random Forest Classifier (RFC) and MapReduce process. The e-health data are collected from the patients who suffered from different diseases is considered for analysis. The optimal attributes are chosen by using Improved Dragonfly Algorithm (IDA) from the database for the better classification. Finally, RFC classifier is used to classify the e-health data with the help of optimal features. It is observed from the implementation results is that the maximum precision of the proposed technique is 94.2%. In order to verify the effectiveness of the proposed method, the different performance measures are analyzed and compared with existing methods.

Keywords Internet of things · Big data · E-health · Map reduce · Random forest classifier · Dragonfly algorithm · Optimization

List of symbols

Sp_i	Separation of ith individual
P	Current position
P_k	Position of kth individual
N	Total number of neighboring individual in the search space
A_{li}	Alignment of ith neighboring individual
V_k	Velocity of kth individual
P^-	Position of enemy
P^+	Position of food source
sw	Separation weight
aw	Alignment weight
cw	Cohesion weight
Att	Attraction, food factor
Dis	Distraction, enemy factor
w_CR	Inertia weight-crossover rate
t	Iteration count
$f_{\max}$	Largest fitness value
f_p	Larger of the two individuals to cross the fitness
f_{avg}	Average fitness
f	Mutation individual's fitness
R_1, R_2	Random values
$V1, V2$	Random vectors that indicate the probability
F	Margin function

✉ Naveen Chilamkurti
N.Chilamkurti@latrobe.edu.au

S. K. Lakshmanaprabu
prabusk.leo@gmail.com

K. Shankar
shankarcrypto@gmail.com

M. Ilayaraja
ilayaraja.m@klu.ac.in

Abdul Wahid Nasir
abdulwahidnasir@bitsathy.ac.in

V. Vijayakumar
vijayakumar.v@vit.ac.in

[1] Department of Electronics and Instrumentation Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

[2] School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, India

[3] Electronics and Instrumentation Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India

[4] School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India

[5] Cyber Security Program Coordinator, Computer Science and IT, La Trobe University, Melbourne, Australia

Published online: 04 January 2019

$I()$	Indicator function
$\arg_k I(h_k(V1)$	h_k is nth tree of the RF

1 Introduction

The IoT refers to the next group of the Internet which will contain trillions of nodes representing different things from sensor devices and handhelds to large web servers and processer clusters [1]. The IoT recognizes and controls things in existing network, clearing a path for chances for the physical world into computer-based systems which can result in enhanced efficiency, precision and economic benefit [2, 3]. In spite of the way that IoT has created unprecedented changes that can help increase revenue, reduce costs, and improve efficiencies, collecting a huge measure of data alone is insufficient [4]. In present days, computers have conveyed huge improvement to technology that leads to the creation of huge volumes of data [5]. Moreover, the advancement of the healthcare database management systems creates a huge number of medical databases [6]. It is a supervised learning approach to grouping model [7].

Data mining is the procedure of breaking down data from the alternative purpose of views and compressing it into useful data [8, 9]. In Big Data, the huge data faces a problem of characterization. When there is high dimensionality of datasets, clustering becomes very moderate [10]. Map Reduce is the method for processing problems including large amounts of data, especially for problems that can easily be assigned into independent subtasks that can be worked out [11]. This absence of data must be taken into account when picking [21] a feature extraction methodology to extract the set of features from an observation will result in the loss of some data [12, 13]. Feature selection methods are well-defined as a subset of the feature extraction field. It chooses features that are relevant to the target ideas [14–16].

The benefit of the feature range includes reducing the data size when unnecessary features are discarded, enhancing the "Classification/prediction precision" [17]. Classification divides data samples into target classes. The Classification technique predicts the target class for each data focuses. For example, the patient can be classified as "high hazard" or "generally safe" patient based on their disease pattern utilizing data characterization approach [18]. The correctness of the classifier could be tested utilizing the test dataset. In the present examination, we have focused on the usage of order systems in the field [19] of medical science and bioinformatics [20, 21].

The Classification is the most generally applied data mining technique and employs a set of pre-classified examples to develop a model that can group the number of inhabitants in records everywhere [22, 23]. One of the attributes, called the ordering attribute [5], indicates the class to which each dataset belongs. The objective of arrangement is the method to assemble a model of the grouping attribute based upon the other attributes which are not from the preparation dataset [22–24].

Main Contribution of this work to classify medical data with an innovative RF classifier with optimal features. This analysis medical health data collected from IoT are useful for medical big data analytics. So, the size reduction model MapReduce framework is considered, and then optimal features are extracted from those data for the classification process. Finally, difference measures are considered to evaluate the performance of the proposed model.

2 Literature review

In 2017 Antunes et al. [25] have proposed the semantic approaches for context association and extend our unsupervised model to learn word categories naturally. The solution was evaluated against "Miller–Charles dataset" and an IoT semantic dataset extracted from a mainstream IoT stage, achieving a correlation of 0.63. Non-negative lattice factorization can likewise be used to discover latent semantic data in distributional profiles and increase precision.

Shadroo and Rahmani [26] investigated the recent research work done on IOT using big data and data mining. An outline of the approaches used in the area of IoT-big data and IoT-data mining is presented in three categories to provide a stage for researchers in the future.

In 2017 Amroun et al. [27]. suggested the best descriptor to recognize human movement utilizing Convolutional Neural Network (CNN). We chose to order four types of activities: standing, sitting, laying and strolling. Results demonstrate that the Discrete Cosine Transform, with the CNN as a classifier, achieves more than 98% average precision in the classification of activities such as standing, sitting, laying and walking.

In 2018 Girish et al. [28] have proposed a supervised audio characterization system utilizing sparse representation over a cloud network. The audio classification is done separately on different nodes utilizing distributed sparse representation [29].

Dragonfly Algorithm (DA) is a recently proposed swarm intelligent algorithm based on the immobile and dynamic swarming behavior of dragonflies [30]. Sree Ranjini and Murugan proposed a method which combines the exploration limit of DA and exploitation capacity of Particle Swarm Optimization (PSO) to achieve global optimal solutions. The efficiency of the DA is validated by testing on fundamental unconstrained and CEC 2014 benchmark problems. A comparative performance investigation between Memory based Hybrid Dragonfly Algorithm (MHDA) and other powerful enhancement algorithms have been carried out

and the significance of the results is proved by verifiable methods. The results demonstrate that MHDA gives better performance than conventional DA and PSO. Moreover, it gives competitive results in terms of convergence, exactness, and search-limit when compared with the state-of-the-craftsmanship algorithms.

Chaudhary et al. in 2016 [23] developed an improved random forest classifier method for multi-class disease order problem. The authors demonstrated that the improved random forest classifier provides better than RFC with an increase in accuracy. It intends to improve the performance of the Random Forest algorithm. The performance results confirm that the proposed improved-RFC approach performs better than Random Forest algorithm with an increase in disease classification exactness up to 97.80% for multi-class groundnut disease dataset.

In the Hadoop framework, Map Reduce is used to perform the cluster filtering, aggregation and to keep up the efficient storage structure. Subramaniyaswamy et al. [32], have proposed an emoticon based clustering method by utilizing sentiment analysis through natural language processing.

In 2016, Singh and Sharma [3] have discussed data mining models in IoT which is a multi-layer model, dispersed model, lattice-based model and big data model. The key issues about the models like a collection of data, data deliberation, and combination, event sifting etc. are likewise discussed. After reviewing the key issues, a new model has been proposed.

In 2018 Yang et al. [33] have proposed the enhanced ID3 algorithm to overcome multi-esteem predisposition issue while choosing test/split traits, illuminates the issue of numeric characteristic discretization and stores the classifier display as standards by utilizing a heuristic methodology for simple comprehension and memory reserve funds. The enhanced ID3 algorithm is better classification as far as precision, soundness and minor error rate.

The missing data in the datasets create large classification errors. The imputation technique is the general approach to transform incomplete data into complete data. Tran et al. [34] have proposed an approach by incorporating imputation and genetic-based feature selection to enhance the speed of classification with incomplete data.

In the existing MapReduce, framework process has a few disadvantages that are precision and classification rate is minimum and also health records in big data, not up to the optimal features with the classification process.

3 Big data and E-health in IoT

The growth and development of Big Data knowledge have opened up a gateway to its applications in e-health, and furthermore is becoming increasingly vital to initiate its integration. The data in IoT has its own characteristics such as constantly quantity, distributed, time-related and position-related. At the same time, the data sources of IoT are varied, and the resources of nodes are limited. For this, different data mining models can be used. The different data mining models for the IoTare a multi-layered model, dispersed model, network-based model and big data model. The use of big data technology in the healthcare is a promising area where it can predict diseases and allocate the patient before health issues which can reduce the cost of treatment and improve the quality of life.

3.1 Application for big data in IoT

- Enormous information advancements can offer information storing and preparing directions in an IoT domain, while information investigation permits businessmen to settle on better choices.
- Most IoT applications do not just spotlight on observing discrete events but in addition on mining the data gathered by IoT information.
- For example, an e-Health application can gauge circulatory strain and glucose levels. A patient utilizing the arrangement has another wellness tracker that can add wellness information to the e-Health arrangement.
- Network arrangements give the adaptability and high accessibility that numerous IoT applications require. The different IoT applications can be implemented in a smart city with various electronics devices [35].

4 Methods

This methodology analyzes e-health big data in IoT with help of some modeling procedures. Most data collection process in the IoT environment are sensor-fitted devices that require custom conventions. The big data are classified according to essential elements that are the size of data, types of data and so on. This analysis of the data's collected from real-time for the testing purpose and internet data are used for the training purpose. The important part of IoT process emphasizes the data collection, monitoring, sharing, computerization, control, and cooperation. After the data collection, the features are extracted from the MapReduce model by using IDA optimization technique. From the optimal features established classify the medical data using RFC. It's offers a variability of information services to users based on knowledge discovery from big data.

4.1 Data collection from IoT

Generally, IoT contains data that can be used to the performance of the systems, infrastructures, and things of IoT

whereas the latter contains data that are the results of the interaction between people, between human and systems, and between systems that can be used to enhance the services provided by IoT. Through the Big Data technologies, all health centers could have access to the information for each patient regardless of where the test was made. Moreover, the tests would be stored in real time, allowing making decisions from the instant that the test has been done to the patient, this graphical representation model is shown in Fig. 1.

Generally, the database contains infrastructures and things of IoT, the interaction between systems are used to enhance the service and these databases contain some nuts and bolts information's of healthcare data such as patient name, age, gender, what kind of disease, what are the medicine is taken and etc.

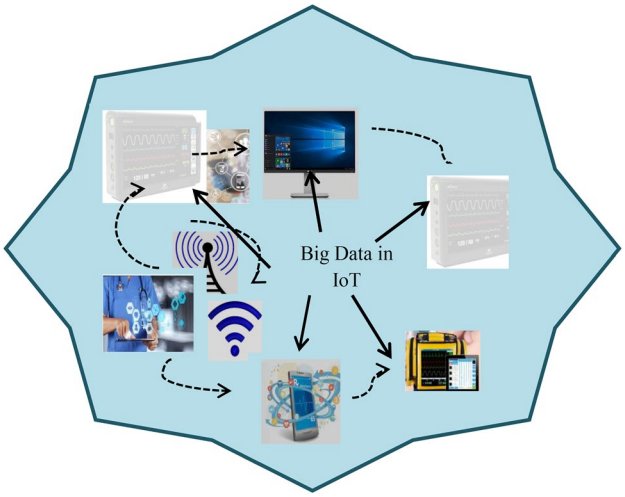


Fig. 2 Graphical model for big data in IoT

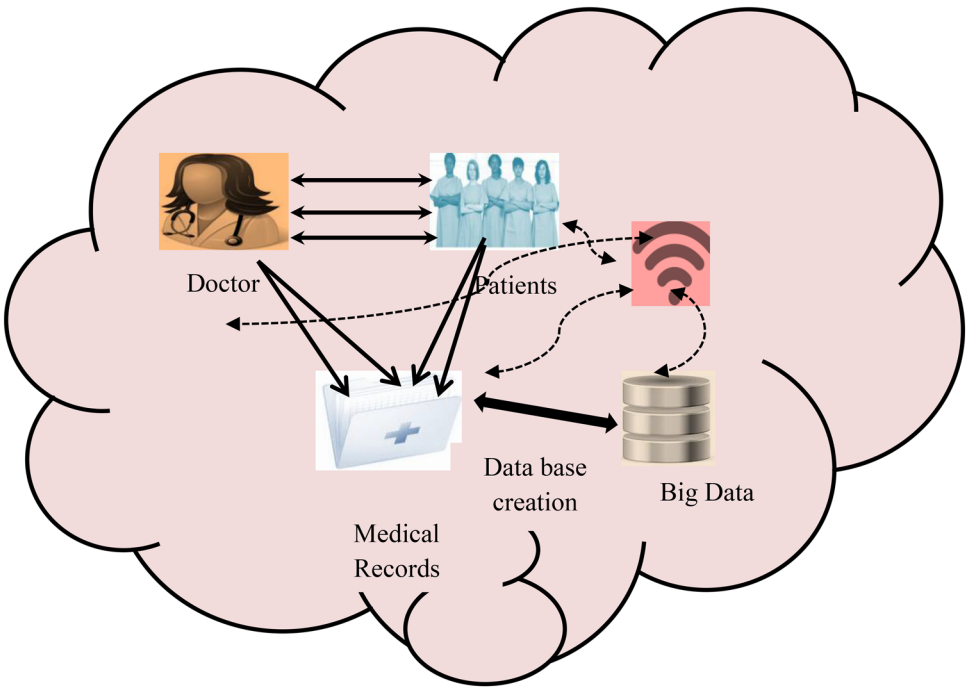
4.2 Data mining model for IoT

The IoT is a very huge and complex structure. Extracting any data from the big data is a troublesome undertaking which can be completed using data mining. For this, different data mining models can be used. E-health big database, initially extracting any data from the smart data is a difficult problem which can be completed using data mining. This process comprises of three procedures for data classifications in IoT which are Hadoop outline work [36], feature extraction to selection and final one as classification model. The graphical model for Big data in IoT is shown in Fig. 2.

4.3 Map-reduce for big data

In the Map Reduce programming model, consists of Mapper, Combiner, and Reducer, which keeps running on all machines in a Hadoop group. The whole map and reduce assignments are executed on various machines in parallel form but the last outcome is gotten simply after the culmination of all the reduce errands. To reduce the size of huge information to this model proficiently processing of enormous information are countless. Here consider gathered health records for Hadoop map-reduce framework process, for the most part is comprised

Fig. 1 Representation for data collection process



of some key parameters which are, Preparation: Prepare the health database, *Map*: Unwanted information expelling and arranging information in light of characters, *Shuffle*: Mix all arranging mapped information, *reduce*: Reducing process performed in each mapped gatherings and last one as gathering all reduced health records or data's.

The info and yield of these capacities must be as (key, value) of information decrease process. Reasonably, a MapReduce undertaking takes input informational index as a key-esteem match and gives yield as key-esteem combine just by processing input informational collections through Map Reduce stages.

4.4 Feature extraction and ranking

Health large information has a few unique features that are not the same as large information from different controls. To evaluate the individual features to examine the information pick up the proportion of each attribute with respect to the class features chose and positioning for advance grouping process. Reduced features are considered the input of the classification model. Here we will utilize Crossover rate based Dragonfly Algorithm, that is Improved DA (IDA) which is used to features selection. once the optimal features are selected to a classification model to classify the healthy or non-healthy data.

4.4.1 Feature selection

This system is utilized to recognize the significant features, which assume a prevailing part in the assignment of the characterization process. The IDA relies upon static and dynamic swarming capacities like partition, arrangement, attachment, fascination towards sustenance source, diversion outwards foes. These two swarming practices are fundamentally the same as the two primary periods of enhancement utilizing meta-heuristics. The attributes were ranked by their information ratio to recognize a class of medical data.

Steps

Initialization Optimal feature selection process initialize the features subsets from the database and also initialize the algorithm parameters, it's described as

$$F_i = \{fe_1, fe_2, \ldots fe_n\}. \tag{1}$$

Separation It refers to the component that individuals take after to maintain a strategic distance from the crash with other neighbors in this separation figured by beneath condition:

$$S_{pi} = \sum_{k=1}^{N} P - P_k. \tag{2}$$

Alignment Next to the separation process, alignment among the dragonflies happens in light of the speed coordinating of individuals to that of different individuals in the neighborhood, it's appeared underneath formula

$$A_{li} = \frac{\sum_{k=1}^{N} V_k}{N}. \tag{3}$$

Cohesion Cohesion eludes to the people's propensity the area's focal point of mass. The union can be figured as:

$$C_{oi} = \frac{\sum_{k=1}^{N} P_k}{N} - P. \tag{4}$$

Attraction and distraction Attraction towards the sustenance source and getting away from foes are other two key practices that every individual carries on to survive. The Distraction outwards a foe between the dragonflies is figured as:

$$Att\_Food_i = P^+ - P \tag{5}$$

$$Dis\_Enemy_i = P^- + P. \tag{6}$$

Updating process DA used two vectors to tackle improvement issues: step vector and position vector, these two vectors are characterized. The progression vector demonstrates the heading of the development of the dragonflies and characterized as takes after:

$$\begin{aligned} \Delta P_{t+1} = &\left(sw \times S_{pi} + aw \times A_{li} + cw \times C_{oi}\right. \\ &\left. + Att\_food_i + Dis\_enemy_i\right) + w\_CR \times \Delta P_t. \end{aligned} \tag{7}$$

Above equations *w_CR* as cross rate weight estimation of hereditary administrators, it's ascertained by utilizing condition (7). And the position vector can be computed as

$$P_{t+1} = P_t + \Delta P_{t+1}. \tag{8}$$

Crossover rate (CR) At last the crossover activity affords the best two arrangements. With the 'n' point hybrid, n cut focuses are randomly picked inside the strings and then fragments between then cut purposes of the two initial solutions are a trade.

$$w\_CR = \begin{cases} R_1(f_{\max} - f_p)/(f_{\max} - f_{avg}), & f_p \geq f_{avg} \\ R_2, & f_p \leq f_{avg}. \end{cases} \tag{9}$$

During the optimization process, different explorative and exploitative practices can be achieved. This leads the merging towards the promising districts of inquiry space and in the meantime, it leads uniqueness outward from the other areas [37] in food search space. In view of this strategy chose attribute information's that is features from e- healthy database.

4.5 Classification model for IoT

Data classification is a process with numerous kinds of existing informational collections for analysis by using the features. Classifying existing IoT-based healthcare network contemplates into three patterns and displaying a rundown of each. This ordered structure is framed in view of the present accessible healthcare arrangements utilizing the IoT help of RFC used to group the information as healthy or not - healthy in light of positioned features.

4.5.1 Random forest classifier

Random Forest is basically a "troupe of unproved classification trees". It gives a fantastic performance on various functional issues, largely because it isn't delicate to the commotion in the informational collection, and it isn't liable to overfitting. It's combining the predictions of several trees, each of which is trained in isolation [31]. RF creates a random example of the information and perceives a key arrangement of ascribes to develop choice trees. The sample structure of RF is shown in Fig. 3.

After the generation of key attributes, it assembles numerous trees and calculated their error rate to choose which tree will be utilized. This order shows inferred by choices tree-based classifiers.

- Every choice tree is developed on an alternate bootstrap test drawn randomly from the preparation information.
- At every node split amid the development of a choice tree, a random subset of m factors is chosen from the first-factor set and the best split in light of these m factors is utilized.

Three main key parameters used in RF classifier which are:

- *Node Size* dissimilar to in decision trees, the number of perceptions in the terminal nodes of each tree of the backwoods can be small. The objective is to develop trees with a minimum bias.
- *Number of Trees* practically 500 trees is regularly a decent decision.
- *Number of Predictors Sampled* the number of predictors tested at each split would appear to be a key tuning parameter that ought to influence how well random forests perform.

For an incomprehensible case, the expectations of the trees that are by the constructed N trees and the classifier blunder esteem shoed by underneath condition.

$$RF_{error} = RF_{r1,r2}(F(V1, V2) < 0). \tag{10}$$

The degree to which the normal number of votes at random vectors for the correct yield surpasses the normal vote in favor of some other yield. The capacity characterized as the

$$F(V1, V2) = \arg_k I(h_k(V1) = V2) - \max \arg_k I(h_k(V2) = j). \tag{11}$$

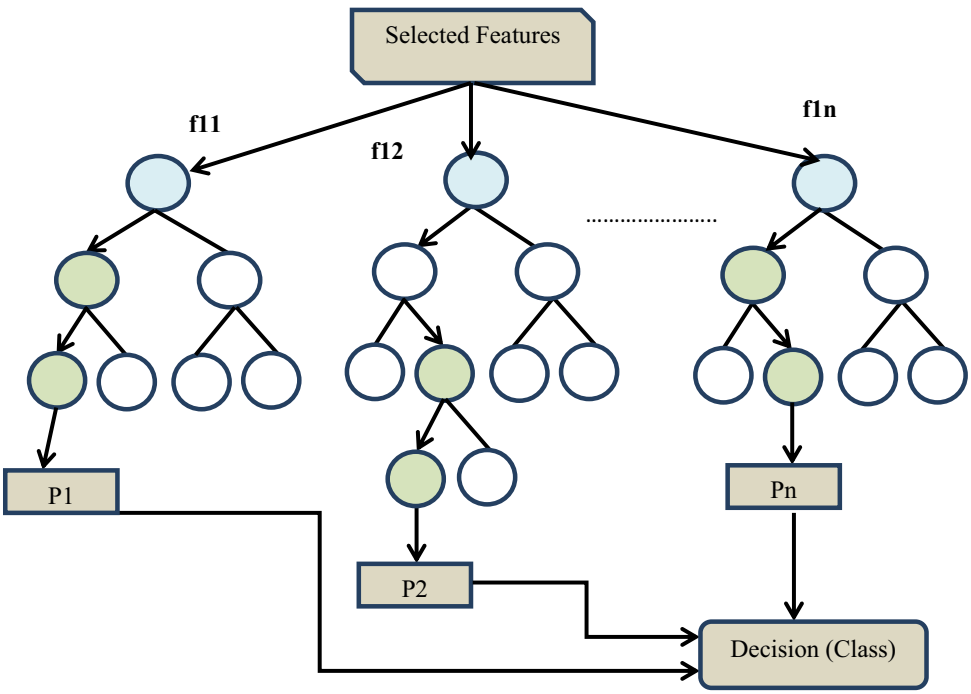


Fig. 3 Sample structure for RF

Two parameters that measure the exactness of individual classifier and reliance between classifiers are quality and relationship, separately. A number of characteristics utilized for base classifier age: this parameter gives the quantity of the number of data is to be utilized which is randomly chosen from the first arrangement of properties at every hub of the base choice tree. This parameter implies the aggregate number of trees utilized as a part of RF demonstrates. The RF show depends on the outcomes decided in each tree uses the most extreme probability technique to acquire ideal outcomes.

Pseudocode for Proposed Model

Input: *Health medical Record Collection*
MapReduce Framework
Reduce Medical database
Feature Selection:
IDA optimization
Initialization
Fitness Evaluation
Condition Checked
If
{
Yes
End
}
Else
{
Separation: Equation (2)
Alignment: Equation (3)
Cohesion: Equation (4)
Attraction and Distraction: Equation (5)
Updating Model: Equation (7)
Improved Process:
Crossover rate (CR): equation (9)
Iter=i+1
}
Ranked Features (Optimal Features)
RF classifier
Output: *Class 0 or 1*

5 Result and discussion

"Our proposed Big data in IoT with e-Health information was executed in java programming with JDK 1.7.0 windows machine containing configurations Intel (R) Core i5 processor, 1.6 GHz, 4 GB RAM" with Hadoopmap-reduce work.

5.1 Database description

Real-time healthcare databases [38–41] of more than 100 K patients per day are considered in this study. In real life, many measuring instruments exist more or less measurement error. Patients influenced by various ailments like coronary illness, kidney sickness, liver maladies and lung issue, this information is gathered from ongoing approved healing facilities. Dataset points of interest are examined in Table 1; it contains a number of patients, characteristic that this features and pertinent class with preparing and testing information.

Table 2, 3 and 4 demonstrate the proposed work the performance investigation of the proposed work and selected features of various databases. If mappers will have changed the performance measures also should be changed. Table 4 shows the validation results of the proposed model, it shows the specificity, specificity, and accuracy results. Validation 1 compared to validation 2 the precision difference is 4.66%, similarly other measures also. The proposed model is a nonspecific approach towards huge information classifying which can be joined with any classifier calculations. Here, sensitivity achieves 92.3%, specificity accomplishes almost 89.52% and 90%, and accuracy acquires over 90% and 95%. In the case of feature extraction, access to the actual data is limited to minimum read actions of the first record of a data frame and only when it is needed to extract the inner schema of a dictionary-based attribute.

Mapper with the gauge reduction database are showed up in Fig. 4, above mentioned four database size is reduced help of Hadoopmap-reduce framework, its diverged from C4.5 techniques to prove our work take minimum time to reduce the size. It's recorded by the second, totally the base time as 52 s for starting mapper.

Figures 5 and 6 exhibit the performance estimations for classifiers and mapper of e-health information gathering process. Precision is portrayed as the amount of the correct forecasts of class names of the informational accumulations in reference to signify instances of the test information. From the analysis combination of mapper and reducer gives better accuracy to predict the class of the dataset. Convolutional Neural Network (CNN), Back Propagation Neural Network (BPNN) and Neural Network (NN) with respect to performance then RFC is the most noteworthy performance. The best precision of this classifier is 94.23% and in perspective of mapper moreover, accuracy is 91.2 similarly sensitivity, specificity and recall also get maximum performance of RFC. As in map reduce model all mappers run parallel so runtime gets reduced. The classifier performance is attempted on getting ready tests from the informational accumulations.

6 Conclusion

IoT works together with big data when voluminous measures of information are ought to have been processed, transformed, and separated in high occurrence. This work,

Table 1 Dataset details

Dataset	Total number of patients	Number of attributes	Class	Training data	Testing data
Heart disease	282	38	2	226	54
Liver disease	180	12	2	147	33
Kidney disease	320	17	2	282	38
Lung disorders	44	42	2	38	6

Table 2 Results of E-health data analysis

Number of mappers	Precision	Recall	F measure	Sensitivity	Specificity
20	92.22	93.56	93.22	79.5	91.2
40	89.22	93.2	98.2	75.55	93.22
60	96.22	90.11	98.2	92.4	83.2
80	94.5	88.6	89.22	79.2	92.14
100	93.33	81.2	79.22	93.2	93.5
120	88.22	79.22	92.2	86.5	94.2
140	90.22	81.22	88.12	88.2	89.56

Table 3 Selections features vs databases

Dataset	DA	IDA	GA
Heart disease	28	34	21
Liver disease	6	9	4
Kidney disease	10	13	9
Lung disorders	32	38	20

we have considered e-health data analysis in IoT with the help of improvement RFC. The IDA is utilized to select the optimal features from the medical dataset which is really helpful for building cost-effective model for disease prediction. The proposed method provides classification accuracy better than the Gaussian mixture model and logistic regression. Both training and testing process get the maximum performance rate that is 94.2% precision, 89.99% recall in the proposed model. The limitation of the proposed algorithm is computationally slow because of the big database. To overcome this issue, the new technique can be developed innovative classifiers and new disease datasets.

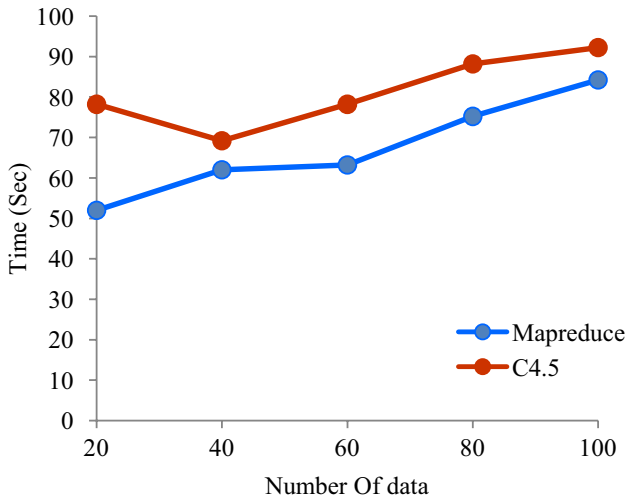


Fig. 4 Mappers vs time analysis

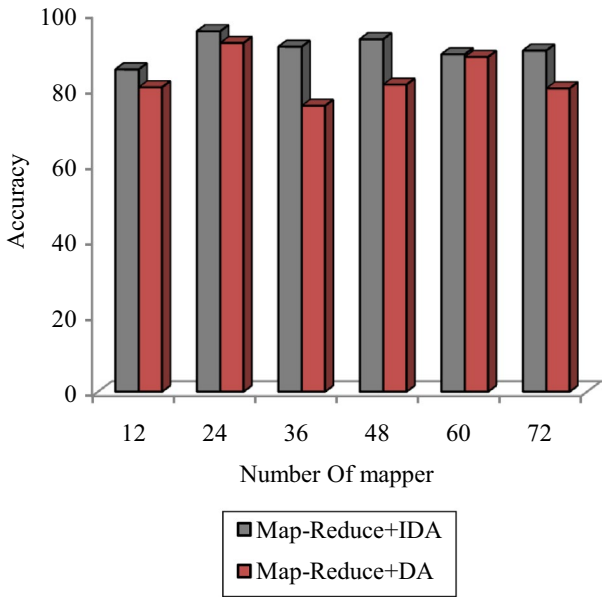


Fig. 5 Accuracy analysis for mappers

Table 4 Validation results

Validations	Accuracy (%)	Recall (%)	F measure (%)	Sensitivity (%)	Specificity (%)
Validation 1	89.4	91.44	94.7	86.23	75.22
Validation 2	93.55	90	91.55	83.12	86
Validation 3	92.22	93.56	90	84.58	80.82
Validation 4	86.33	89.77	92.44	75.22	71.22

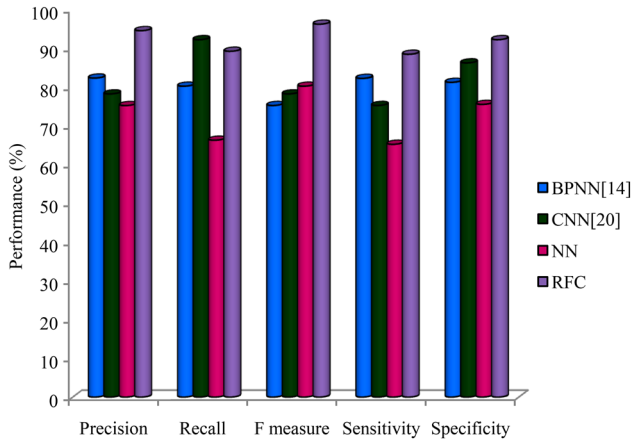


Fig. 6 Comparison of classifiers

References

1. Bin S, Yuan L, Xiaoyi W (2010) Research on data mining models for the internet of things. In: Image analysis and signal processing (IASP), 2010 international conference on, IEEE, pp 127–132
2. Paul A, Daniel A, Ahmad A, Rho S (2017) Cooperative cognitive intelligence for the internet of vehicles. IEEE Syst J 11(3):1249–1258
3. Singh A, Sharma S, 2017, February. Analysis of data mining models for internet of things. In: I-SMAC (IoT in social, mobile, analytics, and cloud) (I-SMAC), 2017 international conference on, IEEE, pp 94–100
4. Yan Z, Liu J, Yang LT, Chawla N (2017) Big data fusion in internet of things. Inf Fusion. https://doi.org/10.1016/j.inffus.2017.04.005
5. Paul A (2013) Graph-based M2M optimization in an IoT environment. In: Proceedings of the 2013 research in adaptive and convergent systems, ACM, pp 45–46
6. Warner JL, Zhang P, Liu J, Alterovitz G (2016) Classification of hospital-acquired complications using temporal clinical information from a large electronic health record. J Biomed Inform 59:209–217
7. Ahmed E, Yaqoob I, Hashem IAT, Khan I, Ahmed AIA, Imran M, Vasilakos AV (2017) The role of big data analytics in the Internet of Things. Comput Netw 129:459–471
8. Plageras AP, Stergiou C, Kokkonis G, Psannis KE, Ishibashi Y, Kim BG, Gupta BB (2017) Efficient large-scale medical data (eHealth Big Data) analytics in the internet of things. In: Business informatics (CBI), 2017 IEEE 19th conference on, IEEE, vol 2, pp 21–27
9. Sugiyarti E, Jasmi KA, Basiron B, Huda M, Shankar K, Maseleno A (2018) Decision support system of scholarship grantee selection using data mining. Int J Pure Appl Math 119(15):2239–2249
10. Susto GA, Schirru A, Pampuri S, McLoone S (2016) Supervised aggregative feature extraction for big data time series regression. IEEE Trans Ind Inform 12(3):1243–1252
11. Masetic Z, Subasi A (2016) Congestive heart failure detection using a random forest classifier. Comput Methods Prog Biomed 130:54–64
12. Revathi L, Appandiraj A (2017) Hadoop based parallel framework for feature subset selection in big data. J Innov Res Sci Eng Technol 4(5):3530–3534
13. Shankar K (2017) Prediction of most risk factors in hepatitis disease using Apriori algorithm. Res J Pharm Biol Chem Sci 8(5):477–484. ISSN 0975-8585
14. Mohapatra C, Rautray SS, Pandey M (2017) Prevention of infectious disease based on big data analytics and map-reduce. In: Electrical, computer and communication technologies (ICECCT), 2017 second international conference on, IEEE, pp 1–4
15. Lakshmanaprabu SK, Shankar K, Khanna A, Gupta D, Rodrigues JJ, Pinheiro PR, De Albuquerque VHC (2018) Effective features to classify big data using social internet of things. IEEE Access 6:24196–24204
16. Shankar K, Lakshmanaprabu SK, Gupta D et al (2018) Optimal feature-based multi-kernel SVM approach for thyroid disease classification. J Super Comput. https://doi.org/10.1007/s11227-018-2469-4
17. Manogaran G, Lopez D, Chilamkurti N (2018) In-Mapper combiner based MapReduce algorithm for processing of big climate data. Future Gener Comput Syst 86:433–445
18. Ke Q, Zhang J, Song H, Wan Y (2018) Big data analytics enabled by feature extraction based on partial independence. Neurocomputing 288:3–10
19. Sindhujaa N, Vanitha CN, Subaira AS (2016) An improved version of big data classification and clustering using graph search technique. Int J Comput Sci Mob Comput 5(2):224–229
20. Wang F, Niu L (2016) An improved BP neural network in the internet of things data classification application research. In: Information technology, networking, electronic, and automation control conference, IEEE, pp 805–808
21. Paul A, Ahmad A, Rathore MM, Jabbar S (2016) Smartbuddy: defining human behaviors using big data analytics in the social internet of things. IEEE Wirel Commun 23(5):68–74
22. Ravichandran K, Nagarasan S (2016) Performance of classification in medical data mining. J Innov Res Comput Commun Eng 4(6):12104–12110
23. Paul A, Rho S (2016) A probabilistic model for M2M in IoT networking and communication. Telecommun Syst 62(1):59–66
24. Sisiaridis D, Markowitch O (2017) Feature extraction and feature selection: reducing data complexity with apache spark. Int J Netw Secur Appl 9(6):39–51
25. Antunes M, Gomes D, Aguiar RL (2018) Towards IoT data classification through semantic features. Future Gener Comput Syst 86:792–798
26. Shadroo S, Rahmani AM (2018) Systematic survey of big data and data mining in the internet of things. Comput Netw 139:19–47
27. Amroun H, Temkit MHH, Ammi M (2017) Best feature for CNN classification of human activity using IOT network. In: The internet of things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData), 2017 IEEE international conference on, IEEE, pp 943–950
28. Girish KV, Ramakrishnan AG, Kumar N (2018(A system for distributed audio classification using sparse representation over cloud for IOT. In: Communication systems & networks (COMSNETS), 2018 10th international conference on, IEEE, pp 342–347
29. Paul A (2014) Real-time power management for embedded M2M using intelligent learning methods. ACM Trans Embed Comput Syst (TECS) 13(5s):148
30. Sree Ranjini KS, Murugan S (2017) Memory-based hybrid dragonfly algorithm for numerical optimization problems. Expert Syst Appl 83:63–78
31. Chaudhary A, Kolhe S, Kamal R (2016) An improved random forest classifier for multi-class classification. Inf Process Agric 3(4):215–222
32. Subramaniyaswamy V, Vijayakumar V, Logesh R, Indragandhi V (2015) Unstructured data analysis on big data using map reduce. Procedia Comput Sci 50:456–465

33. Yang S, Guo JZ, Jin JW (2018) An improved Id3 algorithm for medical data classification. Comput Electr Eng 65:474–487
34. Tran CT, Zhang M, Andreae P, Xue B, Bui LT (2018) An effective and efficient approach to classification with incomplete data. Knowl Based Syst 154:1–16
35. Talari S, Shafie-khah M, Siano P, Loia V, Tommasetti A, Catalão JP (2017) A review of smart cities based on the internet of things concept. Energies 10(4):421
36. Ayma VA, Ferreira RS, Happ P, Oliveira D, Feitosa R, Costa G, Plaza A, Gamba P (2015) Classification algorithms for big data analysis, a map reduce approach. Int Arch Photogramm Remote Sens Spat Inf Sci 40(3):17
37. Harris NL, Jaffe ES, Stein H, Banks PM, Chan JK, Cleary ML, Delsol G, De Wolf-Peeters C, Falini B, Gatter KC, Grogan TM (1994) A revised European–American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. Blood 84(5):1361–1392
38. https://archive.ics.uci.edu/ml/datasets/heart+Disease. Accessed 10 May 2018
39. https://archive.ics.uci.edu/ml/datasets/liver+disorders. Accessed 4 May 2018
40. https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease. Accessed 6 May 2018
41. http://archive.ics.uci.edu/ml/datasets/Lung+Cancer. Accessed 7 May 2018