International Journal of Engineering & Technology, 7 (1.5) (2018) 97-106

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

LDT-MRF: Log decision tree and map reduce framework to clinical big data classification

T. Surekha[1*], R. Siva Rama Prasad [2]

[1]Associate Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, AP, India.
[2]Research Director, Acharya Nagarjuna University, Guntur, AP, India.
**Corresponding author E-mail: tsurekha1234@gmail.com*

Abstract

The growth of the data is enormous in the current scenario of the developing information technology and performing the data classification is complex both in time and information extraction. Moreover, there are uncertainties in performing the big data classification that are associated with the unbalanced datasets. In order to overcome the issues, a novel method of big data classification is introduced in this paper. The novel method, Log Decision Tree and Map Reduce Framework (LDT-MRF) uses the Log Decision Tree (LDT) and the Map Reduce Framework (MRF) for performing the parallel data classification. The novel parameter termed as Log-entropy is used to select the best feature attribute for data classification. The data classification is performed using the LDT that enables the efficient data classification. Experimentation is carried out using three datasets, namely the Cleveland dataset, Switzerland dataset, and the Breast Cancer dataset. The comparative analysis is carried out using the performance metrics, such as sensitivity, specificity, and accuracy to prove the effectiveness of the proposed method. The sensitivity, specificity, and accuracy of the proposed method is 84.7596%, 74.633%, and 80.9088% respectively, which is greater when compared with the existing methods of big data classification.

Keywords: *Big data classification, Map Reduce, Log-entropy, Log Decision Tree, Accuracy.*

1. Introduction

The impact of the developing information technologies have contribute towards a tremendous growth of the data that is stored, processed, produced, shared, analyzed, and visualized. In 2012, IBM [10] took a survey and reported the amount of data created as 1.5 quintillion bytes in a day and in particular, 90% of the data was created in the past two years [1]. Huge datasets are composed together into a single term named as the big data and they hold 4"V" defined by volume, variety, velocity, and value (e.g. medical images, electronic medical records (EMR), biometrics data, etc.) [2]. Big data is a collection of the datasets whose size and complexity is the major challenges of the standard database management systems and causes a huge challenge to the knowledge extraction techniques. The data is collected from a number of sources, such as the sensors, digital pictures and videos, purchase transactions, social media posts, health care, everywhere [12]. The usage of the Electronic Health Records (EHRs) in the healthcare units initiates the analysis of the healthcare data in order to take necessary actions for improving the health of the patients in an efficient way. Moreover, the exploration of the health care or the medical data is the challenging task due to their heterogeneity, incompleteness, unbalanced, and high dimension in nature.

The medical data contains the recordings of the patients, which is more often heterogeneous data with the various types of the values comprising of the real and the integer values of various ranges, image, and text types. In most of the instances, the medical data is not collected for a purpose but the data is the finalized collection of the data from the health care system. The diagnostic tests, data components, and the other related tests are not performed unless they are strictly essential since the tests are very dangerous and cost effective. The main criterion is that the classes of the patients who have diseases are less when compared with those who do not have the disease, which makes data incomplete and unbalanced. Moreover, the analysis and the knowledge extraction of the data encourage the fact that when more data is available, the information derived from the data is very precise. Thus, it is not possible for the standard algorithms to handle the huge datasets [13] that poses the need to model and adapt the classification algorithms that considers the solutions used in big data. This ensures the proper predicting capacity [1] and hence, the demand for the advanced data driven and machine learning techniques are increasing day-by-day [4].

On contrary, the techniques that deal the big data require the fast, scalable, and parallel implementations, which are satisfied by the Map-Reduce procedure [14]. Map Reduce procedure divides the original dataset and forms the subset that ensures easy handling and finally, all the partial solutions are combined. There are two computational steps in the Map Reduce programming, namely Map and Reduce. Among the two steps, initially, in the map phase, the input data set is subdivided into number of independent problems using the master nodes and then, distributed to the worker independent nodes.

The worker nodes work parallel and compute the sub-problems and return the value to the master nodes. The solutions of the subproblems are then together by the master node to form the output [3] in the reduce phase. Hence, the map reduce is defined based on the required data structure known as <key, value> pair, and the processed data, the intermediate, and the final results follow this data structure. The function of the map and reduce are seen in the following way. The input to the map function is the

<key, value> pair and creates an intermediate <key, value> pair as output, which is further ordered and shuffled based on the intermediate key. This will be the input to the reduce function and generates the required <key, value> pair as final output of the algorithm [9].

This paper proposes an efficient framework for performing the big data classification using the MRF that employs the LDT algorithm for developing the decision tree. The MRF is an effective framework that is capable of performing the parallel processing of the big data for which it employs two functions, namely the map and the reduce function. The sub-sets of the big data are formed using the master node of the MRF that enables parallel processing and the master nodes distributes the sub-sets to the worker nodes that holds the mappers. The map function is present in the mapper that follows the LDT algorithm for developing the LDT model of the sub-set of the data loaded from the big data. Thus, each mapper develops a LDT model that is presented to the aggregator that extracts the data from the model. The extracted data are combined to form the gross data and the gross data is finally developed into a gross LDT model. The gross LDT model is used as a key for generating the class value of the newly arrived big data in the testing phase. The proposed method serves as an easy and efficient method for performing the big data classification.

The contribution of the paper is presented below:

LDT enabled Map-reduce: The Map-reduce framework uses the LDT algorithm to develop the LDT model for the subsets of data loading from the big data.

The paper is organized as: Section 1 introduces the paper, section 2 presents a motivation of the paper presenting the existing methods of big data classification along with their drawbacks. IN section 3, the LDT model is discussed that discusses the procedure. The the proposed method of big data classification using the LDT-MRF is presented with the algorithmic steps in section 4 and section 5 presents the results and discussion of the proposed method that highlights the superior performance of the proposed method. Finally, section 6 concludes the paper.

2. Motivation

2.1 Related works

Victoria López *et al.* [1] proposed a method of big data classification using the Chi-FRBCS-Big DataCS algorithm, a fuzzy rule based classification system that handled uncertainty. The algorithm used the map reduce framework for distributing the problems of the fuzzy model, which is a Cost-sensitive linguistic fuzzy rule-based classification systems. The performance of the method was found to be better and the method handles the imbalanced big data. The computational time was effective but there are small problems associated with the data intrinsic problems that include the small sample size problem. Emad A Mohammed *et al.* [2] reviewed the applications of the Map Reduce programming framework and its implementation platform Hadoop using the clinical data. The main aim was about the ways of enhancing the results of the clinical big data analytics tools. This framework was found to improve the performance of common signal detection algorithms for pharmacovigilance. But it failed to handle the highly uncertain data. Sara del Río *et al.* [3] analyzed the performance of several techniques employed for handling the imbalanced datasets in the big data scenario for which a Random Forest classifier is used. Specifically, oversampling, under sampling and cost-sensitive learning were adapted to big data using Map Reduce that sharply identified the underrepresented class. The main advantage is that the Random Forest classifier provides a good classification platform due to its performance, robustness and versatility. However, large number of the mappers affects the performance, due to the small sample size. Shamsul Huda *et al.* [4] aimed at achieving a fast, affordable and objective diagnosis of the genetic variant of oligodendroglioma with a novel data mining approach that combined a feature selection and ensemble-based classification. In order to reduce the effect of an imbalanced healthcare dataset, a global optimization based hybrid wrapper-filter feature selection was used along with ensemble classification. Though the method handled the imbalanced data, the computational cost was very high. Magnus Orn Ulfarsson *et al.* [5] proposed a classification method based on linear discriminate analysis (LDA) that estimated the covariance using a sparse version of noisy principal component analysis (nPCA). The application of sparsity aimed at the selection of the relevant variables for the classification. This method was even able to handle the microarray of data but the SIS of the gray matter was affected. Alberto Fernández *et al.* [6] proposed a method to handle the scalability issues of the traditional learning approaches for which a Map Reduce framework was used as a “de facto” solution. The classification method was very accurate even when the full dataset is used but the classification model suffered from the problem of dimensionality. Dawen Xia *et al.* [7] used a Map Reduce-Based Nearest Neighbor Approach for predicting the flow of the Big-Data-Driven Traffic. The scalability and the efficiency of the traffic flow prediction were improved but the computational complexity was very high. Sina Khanmohammadia and Chun-An Choua [8] proposed a Gaussian Mixture Model based Discretization Algorithm (GMBD) to preserve the most frequent patterns of the original dataset by considering the multimodal distribution of the numerical variables. Six different publicly available medical datasets were used to analyze the effectiveness of the proposed algorithm and the advantage is that the proposed method fits any domain in the continuous format but the disadvantage is that the computation process is expensive. Sara del Rio *et al.* [9] proposed the Chi-FRBCS-Big Data algorithm, a linguistic fuzzy rule-based classification system along with the Map Reduce framework to learn and fuse rule bases. This algorithm handled the big collections of data with good accuracy and fast response time, but this version yields slower models.

2.2 Challenges

- The major challenge of handling the big data is regarding the design and development of the scalable and parallel algorithms to recognize and classify the patterns.
- The issue existing in the clinical data is regarding the missing information and not the missing data that supports the clinical decision-making [16].
- IN [1], the big data classification method uses the traditional fuzzy classifier to classify the samples of the data, which is based on the fuzzy rule inference. Using fuzzy rules for the big data classification does not yield better results of data classification.
- Most of the traditional classification algorithms used for classifying the big data performs the classification at a very high cost and the high scalability when compared to the map reduce framework.
- The most important challenge of designing the map and the reduce tasks for the proper big data classification relies on the classification algorithm and the success of the map reduce framework depends on the classification algorithm utilized.
- The unbalanced datasets suffer from uncertainties that result in the complex and unbalanced data classification.

3. LDT: -LOG decision tree

The LDT model [27] is the classification model used for the generation of the knowledge from the clinical data that is interpretable and enables easy diagnosis of the diseases. The method is advantageous because of the superior classification accuracy and interpretability when compared with the other methods. The LDT model is the classification model based on the feature attributes of the clinical data and it uses the log-entropy for the selection of the features. The advantage of the feature selection is that it enables the dimensional reduction through the selection of the highly important feature leading to the effective knowledge discovery. The log-entropy of all the features present in the data is computed and the feature attribute with the maximum value of the

log-entropy is selected as the best feature. The log-entropy is based on the weighing function of the features and the entropy of the features. Once the feature is selected, the splitting criterion is continued that performs the optimal calculation of the split point based on the log information gain. Thus, the LDT model is developed and stored that is inferred for the future references.

4. LDT-MRF: A novel LDT algorithm for big data classification using the mapreduce framework

Map Reduce is the programming model that is used to handle the big data and it is used in variety of applications. It consists of two significant tasks, namely the Map and Reduce. The main advantage of using the MRF [9] for the big data classification is that the MRF possess the capacity to process the data at the multiple computing nodes and possess a good scalability. The Map function takes the sub-sets of data loading from the big data and converts each sub-set as a LDT. The construction of the LDT from the data follows some decision rules and they are based on the log-entropy based function for developing the decision tree. Thus, the individual map function generates an individual LDT that is later assembled using the reduce function in the aggregator. The function performed by the aggregator is that it accepts the LDT generated from the individual map function and generates the data from those LDT's. The data is generated from the LDT based on the preset limit and the data generated is less when compared with the original data. The generated data using the LDT is then combined to form a gross data, which is then employed to develop a gross LDT. The gross data is used as a key in the mapper in the testing phase. When the test data arrives, the big data is distributed to the mappers that generate the class values of the subsets of the data loaded from the big data. The function of the aggregator is to generate the unique class value and the classification follows the output of the training phase. Figure 1 depicts the classification model of the MRF.

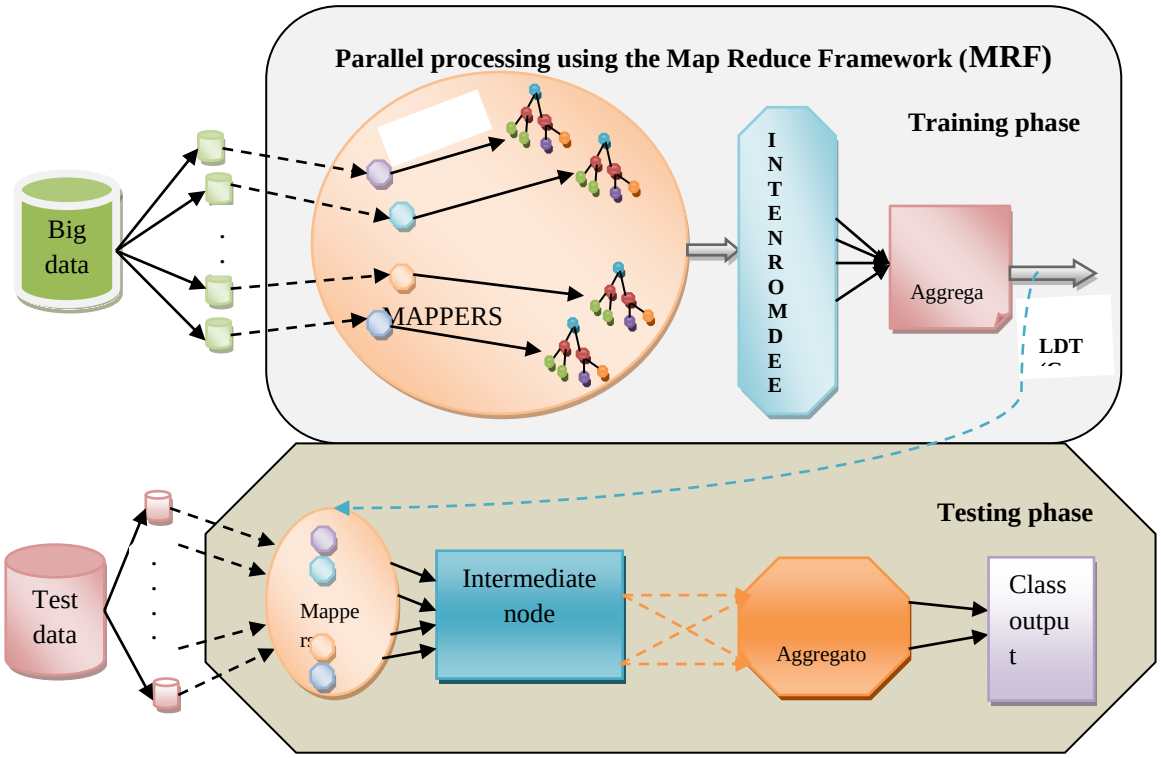


Fig. 1: MRF for the big data classification using the LDT algorithm.

4.1 Reading input data

The input to the MRF is the big data for which the clinical big data is provided. The clinical big data is classified into subsets as the framework is not capable of running the big data as it leads to complexity. Since the MRF supports parallelism, the subsets of data are loaded in the MRF mappers that use the LDF algorithm for the big data classification. Let us consider the big data, which is represented as, D_k .

$$D_k = \{D_1, D_2, ..., D_g\} \quad (1)$$

where, g is the total number of data present in the big data. The dimension of the big data is represented as $(p \times q)$. The subsets are generated using the master nodes of the MRF and the sub-sets are fed to the worker nodes of the MRF such that the worker nodes contain the mappers built with the LDF model. The sub-sets of the data loading from the big data D_k is represented as,

$$D_s = \{d_1, d_2, ...d_r, ..., d_n\} \quad (2)$$

The big data possess the data attributes and the data attributes present in the big data is given by,

$$f_M = \{f_1, f_2, ..., f_h, ... f_x\} \quad (3)$$

Where, f_M is the data attributes of the big data and X is the total number of attributes present in the big data.

4.2 Training phase

In the training phase, the key LDT is generated from the gross data, which is named as the gross LDT. The MRF consists of the mappers and the aggregators functioning with the map functions map and reduce.

The mappers and aggregators are the present in the worker nodes that processes the subsets of the data loading from the big data provided by the master node.

A. Mapper phase

In the mapper phase, the subsets of data are fed to the worker nodes that are uploaded with the LDT model. In this phase, the

LDT models for all the subsets of the data are generated and the total number of the generated LDT models is n. Let us consider there are n number of worker nodes with the map function and the number of mappers present in the framework is denoted as,

$$m^{'} = \{m_1, m_2, ..., m_r, ..., m_n\} \quad (4)$$

where, n is the total number of mappers present in the MRF during the training phase. The mappers are built with the LDT algorithm that accepts the sub-sets of data loaded from the big data and generates individual LDT model for each big data sub-sets and the Mapper loaded with the sub-sets of big data is represented as, $m_r(d_r)$ and it generates the output equal to $LDT_r(d_r)$.

where, $m_r(d_r)$ represents the r^{th} mapper operating with the input d_r and $LDT_r(d_r)$ corresponds to the LDT model of the data sub-set d_r loaded from the big data D_k.

B. Intermediate data

The intermediate nodes collect the n number of the LDT models generated using the worker nodes present in the mapper phase of the MRF. The output from the mapper is represented as,

$$M^o = \{LDT^1, LDT^2, ..., LDT^r, ..., LDT^n\} \quad (5)$$

$$M^r = [LDT^r]; r \le n \quad (6)$$

where, M^o is the output from the mapper, and M^r is the output of the r^{th} mapper. n is the total number of the mappers present in the framework, each constituting to the individual LDT. Therefore, n number of LDT models is obtained.

$$[LDT^G] = R[LDT^1, LDT^2, ..., LDT^r]; 1 \le r \le n \quad (7)$$

where, $LDT^1, LDT^2, ..., LDT^r$ are the LDT models of the individual data subsets loading from the big data. LDT^G represent the LDT model of the gross data. These LDT's generated using the mapper is applied to the aggregator through a intermediate node. The aggregator accepts all the LDT corresponding to the subsets of the data and extracts the data using the LDT models.

C. Reducer phase

The data extracted using the LDT models are combined to form the gross data for which the LDT algorithm is applied to develop a gross LDT.

$$LDT^G = LDT(G^{'}) \quad (8)$$

where, LDT^G is the decision tree output of the aggregator and the gross data is represented as,

$$G^{'} = [d_1^* \| d_2^* \| .... d_r^* \| ... d_n^*] \quad (9)$$

$[d_1^* \| d_2^* \| .... d_r^* \| ... d_n^*]$ Represents the gross data obtained from the $LDT^1, LDT^2, ..., LDT^r, ..., LDT^n$. $G^{'}$ is the gross data generated from the LDT models.

4.3 Testing phase

The main aim of the testing phase is to generate the class value of the arriving data or in other words, the clinical data is classified and the corresponding class value is presented in the output of the testing phase. When a new test data arrives at the MRF, the class value of the data is generated using the key gross LDT in order to attain an efficient classification. Let us represent the test data as,

$$D_Q^t = \{d_1^t, d_2^t, ..., d_N^t, ... d_s^t\} \quad (10)$$

where, $d_1^t, d_2^t, ..., d_N^t, ... d_s^t$ are the subsets of the test big data loading from D_Q^t.

A. Mapper phase

The worker nodes are provided with the map function that follows the gross LDT for generating the class value. Each of the mappers generates the class value and the number of the class equals to the number of mappers present in the MRF during the testing phase. The number of mappers engaged in the testing phase is denoted as,

$$m^t = \{m_1^*, m_2^*, ..., m_N^*, ... m_s^*\} \quad (11)$$

The mapper output is represented as, $m_N^t(d_N, LDT^G)$ that holds the gross LDT and the sub-sets of the test data. The mapper in the testing phase is provided with the reference LDT^G that is obtained from the aggregator of the testing phase. The testing mapper generates the class values and the class values generated in the testing phase is represented as,

$$C = \{C_1, C_2, ..., C_N, ..., C_s\} \quad (12)$$

B. Intermediate data

The intermediate data present in the intermediate node is the class values of the test data loaded in the MRF from the test database. The class values of the test data is provided in the equation (12) and the class values are generated in the mapper and fed to the intermediate node for aggregating the class values in the aggregator.

C. Reducer phase

The aggregator accepts the class values generated from all the mappers and produces the respective class value. For illustration, if there are four mappers and they yield four class values then, the function of the aggregator is to combine all the class one value and produce a single class one output. Similarly, the other class outputs are generated yielding four class values. The output from the reducer is denoted as,

$$O^N = R(C_N) \quad (13)$$

where, $1 \le N \le s$ and S is the total number of the class output. The class output is used for the efficient big data classification that reduces the complexity and thus reduces the processing time.

5. Results and Discussion

In this section, the results and discussion is presented to prove the importance of the proposed LDT+MRF when compared with the existing methods and the superiority is proved based on the performance metrics, namely sensitivity, specificity, and accuracy.

5.1 Experimental setup

The experimentation is carried out using three medical datasets such as, Cleveland, Switzerland and Breast Cancer data available in the UCI machine learning repository [26]. The experimentation

is performed in Windows 8, 4GB RAM and the implementation is carried out in JAVA programming with map reduce libraries.

5.2 Evaluation metrics

The performance of the proposed LDT-MRF classifier will be analyzed using sensitivity, specificity and accuracy.

5.2.1 Sensitivity

Sensitivity is the proportionality existing between the false negative and the true positive.

$$Sensitivity = \frac{1}{1+\frac{FN}{TP}}$$

5.2.2 Specificity

Specificity denotes the proportionality between the false positive and the true negative.

$$Sensitivity = \frac{1}{1+\frac{FP}{TN}}$$

5.2.3 Accuracy

Accuracy shows the proportionality between the false positives and the true positives.

$$Accuracy = \frac{1}{1+\frac{FP}{TP}}$$

5.3 Methods taken for comparison

The proposed LDT-MRF classifier will be compared with the existing algorithms to prove the performance improvement of the proposed algorithm. The methods taken for comparison include: DT+MRF, HDT+MRF, ANN, and the proposed LDT+MRF.

5.4 Comparative analysis based on the evaluation metrics – sensitivity, specificity, and accuracy

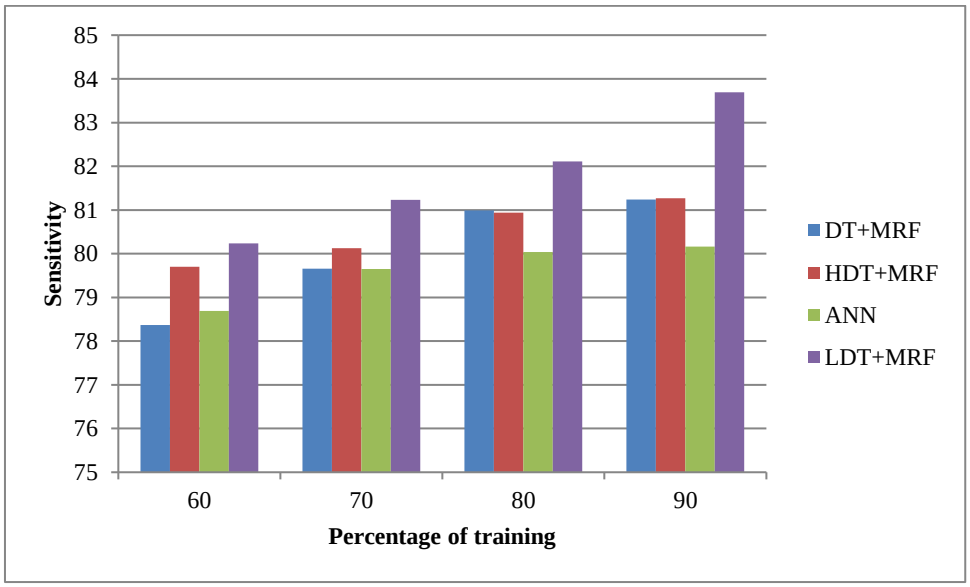
In this section, the comparative analysis of the proposed LDT+MRF is carried out in terms of the performance metrics, such as sensitivity, specificity, and accuracy.

5.4.1 Using the Cleveland dataset

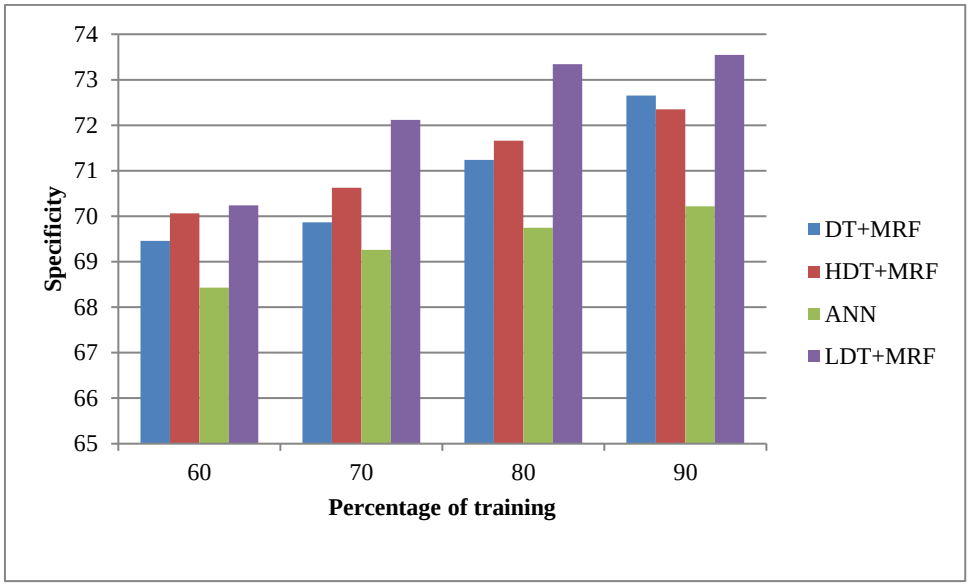
The Cleveland dataset is utilized to perform the comparative analysis of the proposed method in terms of the specificity, sensitivity, and accuracy. Figure 2 a) shows the comparative analysis of the Cleveland dataset based on sensitivity. When the training percentage is 60, the percentage of sensitivity attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 78.36584, 79.6987, 78.694, and 80.2369 respectively. The sensitivity of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 81.23658%, 81.2698%, 80.1635%, and 83.6948% respectively. It is very clear from the values that the percentage of sensitivity is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of sensitivity for the proposed method increases from 80.2369% to 83.6948% and the percentage of sensitivity increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of sensitivity.

Figure 2 b) shows the comparative analysis of the Cleveland dataset based on specificity. When the training percentage is 60, the percentage of specificity attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 69.458, 70.063, 68.427, and 70.237 respectively. The specificity of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 72.653%, 72.348%, 70.218%, and 73.546% respectively when the percentage of training is 90. It is very clear from the values that the percentage of specificity is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of specificity for the proposed method increases from 70.237% to 73.546% and the percentage of specificity increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of specificity.

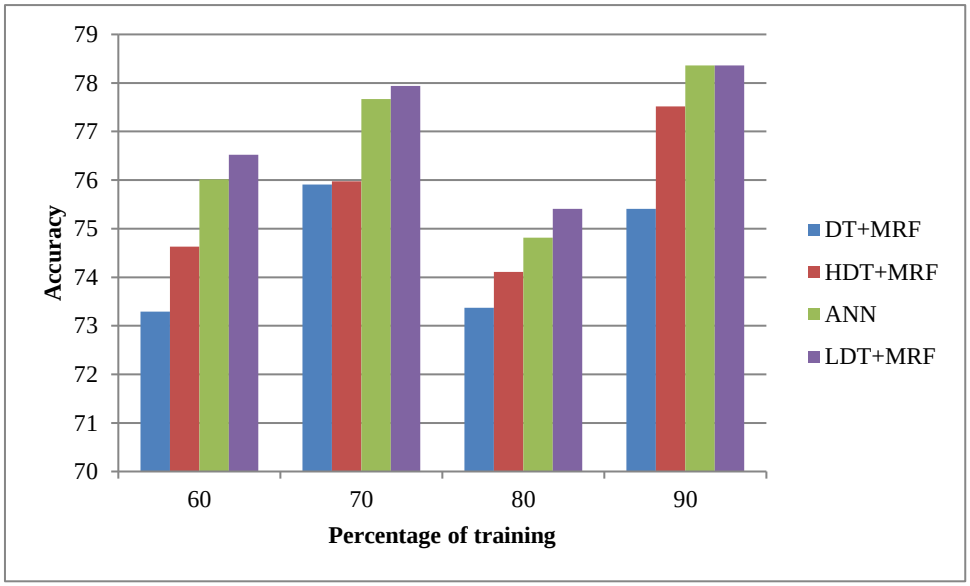
Figure 4 c) shows the comparative analysis of the Cleveland dataset based on accuracy. When the training percentage is 60, the percentage of accuracy attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 73.2939, 75.9087, 73.3713, and 75.4087 respectively. The accuracy of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 76.5234%, 77.9412%, 75.4077%, and 78.3641% respectively when the percentage of training is 90. It is very clear from the values that the percentage of accuracy is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of accuracy for the proposed method increases from 75.4087% to 78.3641% and the percentage of accuracy increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of accuracy.



a) Sensitivity of the methods using the Cleveland dataset



b) Specificity of the methods using the Cleveland dataset



c) Accuracy of the methods using the Cleveland dataset

Fig. 2: Comparative analysis using the Cleveland dataset.

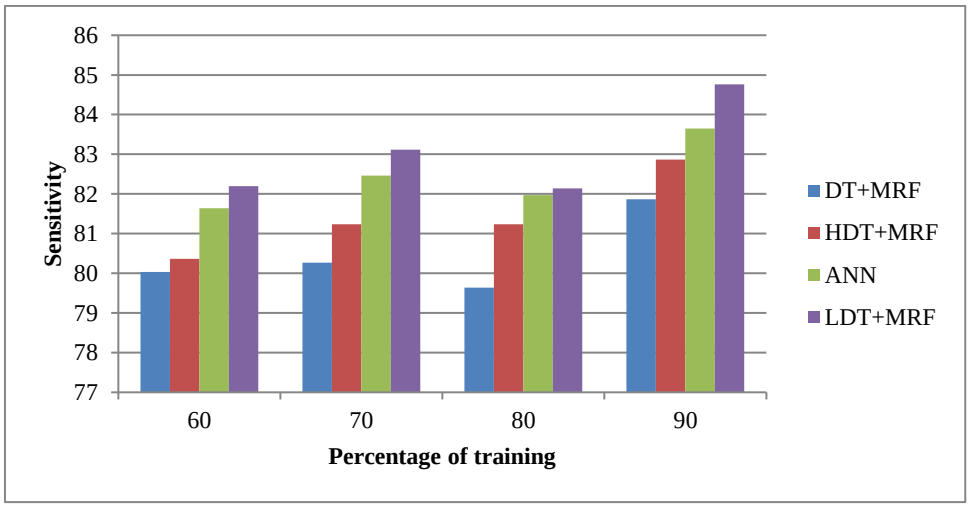
5.4.2 Using the Switzerland dataset

The Switzerland dataset is utilized to perform the comparative analysis of the proposed method in terms of the specificity, sensitivity, and accuracy. Figure 3 a) shows the comparative analysis of the Switzerland dataset based on sensitivity. When the training percentage is 60, the percentage of sensitivity attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 80.0365, 80.2697, 79.639, and 81.8634 respectively. The sensitivity of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 81.6397%, 82.463%, 81.9678%, and 83.648% respectively when the training percentage is 80. It is very clear from the values that the percentage of sensitivity is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of sensitivity for the proposed method increases from 81.8634% to 84.7596% and the percentage of sensitivity increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of sensitivity.

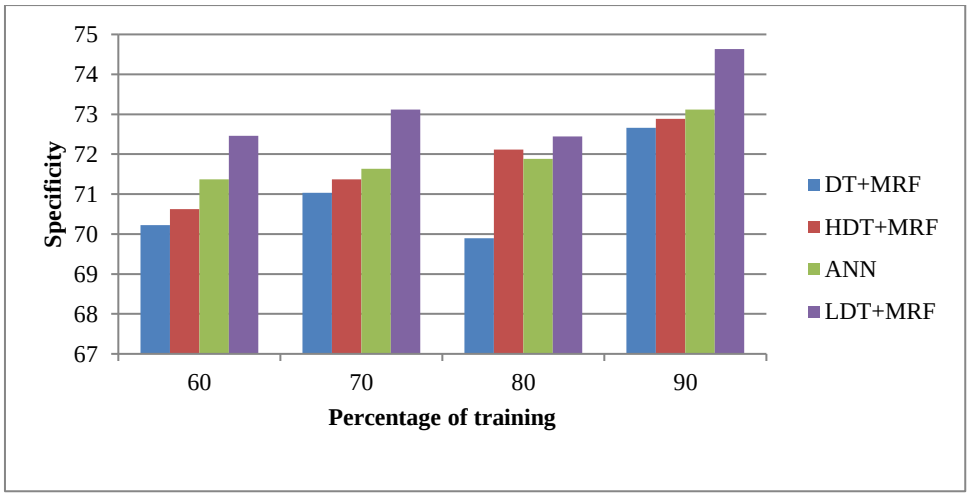
Figure 3 b) shows the comparative analysis of the Switzerland dataset based on specificity. When the training percentage is 60, the percentage of specificity attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 70.224, 71.037, 69.896, and 72.664 respectively. The specificity of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 72.461%, 73.115%, 72.446%, and 74.633% respectively when the percentage of training is 90. It is very clear from the values that the percentage of specificity is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of specificity for the proposed method increases from 72.664% to 74.633% and the percentage of specificity increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of specificity.

Figure 3 c) shows the comparative analysis of the Switzerland dataset based on accuracy. When the training percentage is 60, the percentage of accuracy attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 75.7201, 76.2973, 74.8766, and 77.5536 respectively. The accuracy of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 78.8858%,

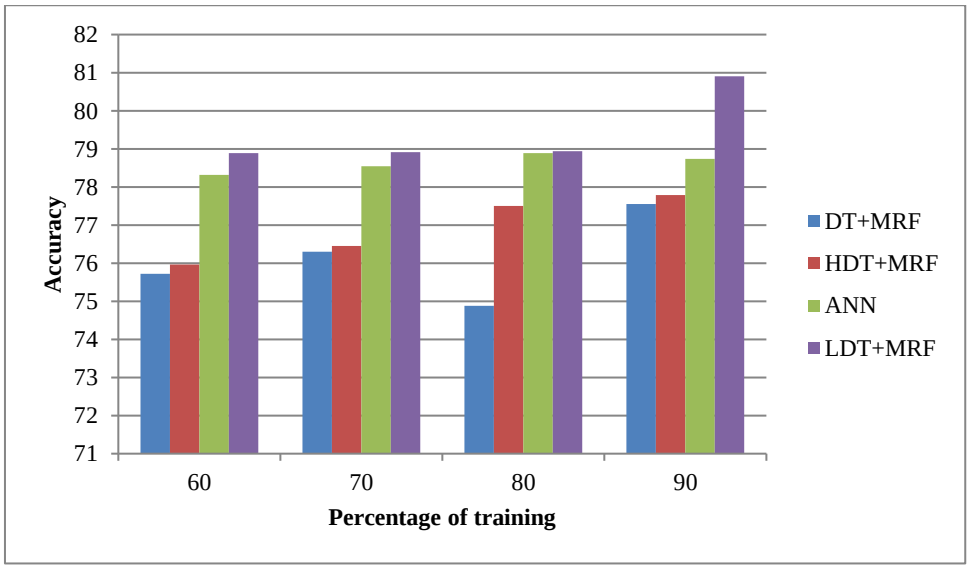
78.9135%, 78.9354%, and 80.9088% respectively when the percentage of training is 90. It is very clear from the values that the percentage of accuracy is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of accuracy for the proposed method increases from 77.5536 % to 80.9088%, and the percentage of accuracy is found to increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of accuracy.



a) Sensitivity of the methods using the Switzerland dataset



b) Specificity of the methods using the Switzerland dataset



c) Accuracy of the methods using the Switzerland dataset

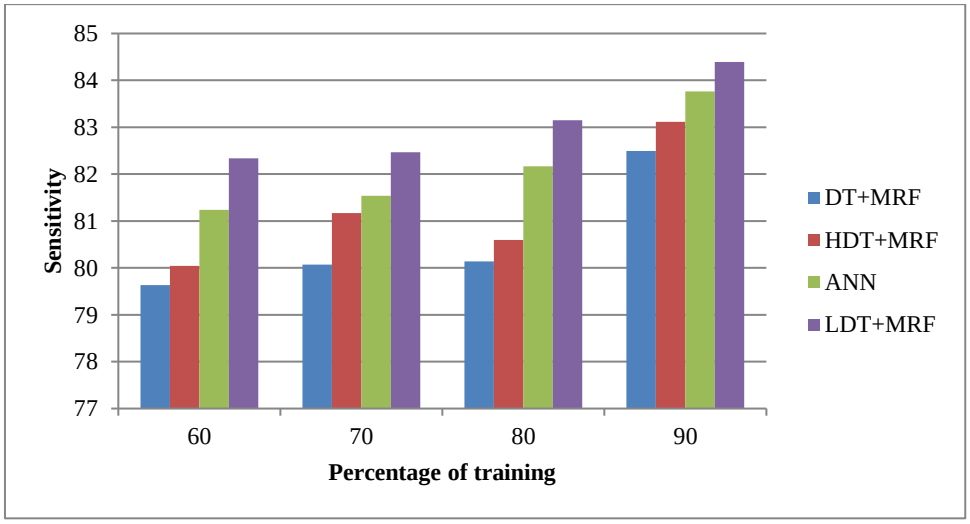
Fig. 3: Comparative analysis using the Switzerland dataset.

5.4.3 Using the Breast Cancer dataset

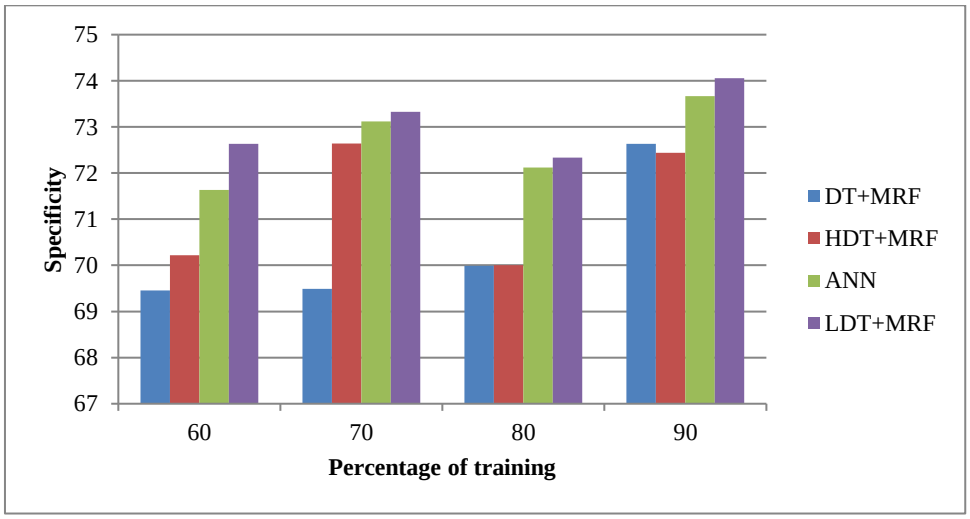
The Breast Cancer dataset is utilized to perform the comparative analysis of the proposed method in terms of the specificity, sensitivity, and accuracy. Figure 4 a) shows the comparative analysis of the Breast Cancer dataset based on sensitivity. When the training percentage is 70, the percentage of sensitivity attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 80.045, 81.1697, 80.598, and 83.116 respectively. The sensitivity of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 82.334%, 82.4698%, 83.1475%, and 84.395% respectively when the training percentage is 80. It is very clear from the values that the percentage of sensitivity is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of sensitivity for the proposed method increases from 82.496% to 84.395%, and the percentage of sensitivity is found to increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of sensitivity.

Figure 4 b) shows the comparative analysis of the Switzerland dataset based on specificity. When the training percentage is 60, the percentage of specificity attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 69.452, 69.488, 69.991, and 72.634 respectively. The specificity of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 72.633%, 73.328%, 72.334%, and 74.054% respectively when the percentage of training is 90. It is very clear from the values that the percentage of specificity is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of specificity for the proposed method increases from 72.634 % to 74.054%, and the percentage of specificity is found to increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of specificity.

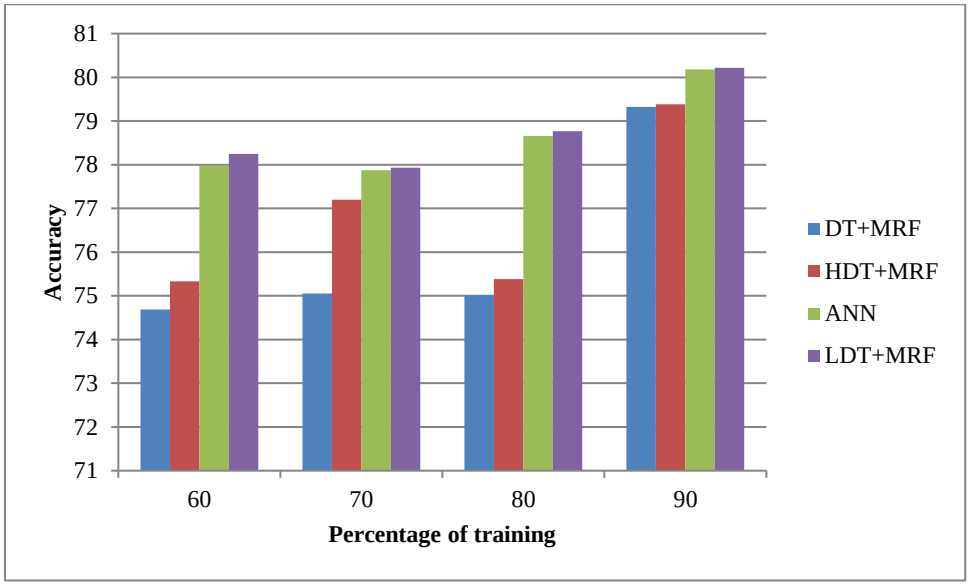
Figure 4 c) shows the comparative analysis of the Switzerland dataset based on accuracy. When the training percentage is 60, the percentage of accuracy attained using the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 74.6853, 75.0521, 75.0195, and 79.3249 respectively. The accuracy of the methods DT+MRF, HDT+MRF, ANN, and LDT+MRF are 78.24935%, 77.9286%, 78.76940%, and 80.2187% respectively when the percentage of training is 90. It is very clear from the values that the percentage of accuracy is greater for the proposed LDT+MRF when compared with the existing methods. Moreover, when the percentage of training increases from 60% to 90%, the percentage of accuracy for the proposed method increases from 79.3249% to 80.2187%, and the percentage of accuracy is found to increase for the other existing methods. Thus, the proposed LDT+MRF outperform the existing methods in terms of accuracy.



a) Sensitivity of the methods using the Breast Cancer dataset



c) Specificity of the methods using the Breast Cancer dataset



c) Accuracy of the methods using the Breast Cancer dataset

Fig. 4: Comparative analysis using the Breast Cancer dataset

5.5 Discussion

Table 1 shows the discussion of the comparative methods in terms of the metrics, namely the sensitivity, specificity, and accuracy using three datasets. The datasets used for the analysis include the Cleveland dataset, Switzerland dataset, and the breast cancer dataset. The proposed LDT+MRF attained a sensitivity percentage of 83.6948**%** for the Cleveland dataset, 84.7596**%** for the Switzerland dataset, and 84.395**%** for the breast cancer dataset. The sensitivity percentage of all the existing methods is less when compared with the proposed LDT+MRF. Similarly, the specificity attained using the proposed method is 73.546**%** for the Cleveland dataset, 74.633**%** for the Switzerland dataset, and 74.054**%** for the breast cancer dataset. The specificity percentage of the proposed method is greater than the ANN, HDT+MRF, and DT+MRF methods. The accuracy of the proposed method attained is 78.3641%, 80.9088%, and 80.2187% respectively for the Cleveland, Switzerland, and Breast Cancer datasets. The table proves that the proposed method attained a greater percentage of sensitivity, specificity, and accuracy.

Table 1: Discussion of the comparative methods

Methods	Cleveland dataset			Switzerland dataset			Breast Cancer dataset		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Proposed LDT+MRF	**83.6948**	**73.546**	**78.3641**	**84.7596**	**74.633**	**80.9088**	**84.395**	**74.054**	**80.2187**
ANN	80.1635	70.218	75.4077	82.1369	72.446	78.9354	83.1475	72.334	78.7694
HDT+MRF	81.2698	72.348	77.9412	83.1178	73.115	78.9135	82.4698	73.328	77.9286
DT+MRF	81.23658	72.653	76.5234	82.1984	72.461	78.8858	82.334	72.633	78.2493

6. Conclusion

The paper presented a novel method of big data classification using the LDT enabled MapReduce framework. The use of the MapReduce framework introduces two function mapping function and the reduce function to perform the classification. These functions are assigned with the LDT structure that formulates the log-entropy model of the big data. Initially, the big data is classified as subsets of big data and are loaded in the mapper that uses the LDT for generating the model. The LDT uses a novel log-entropy parameter to classify the data and the model generated is combined in the aggregator to form a gross LDT for which the generated LDT using the mapper is converted into the gross data. Finally, when the new data arrives for classification, the gross LDT is utilized by the mapper to generate the class label to perform classification. The experimentation carried out using the three datasets prove that the proposed method of big data classification attained a greater sensitivity, specificity, and accuracy of 84.7596%, 74.633%, and 80.9088% respectively that is better compared with the existing methods like the ANN, HDT+MRF, and DT+MRF.

References

[1] Victoria López, Sara del Río, José Manuel Benítez, and Francisco Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the Map Reduce framework for imbalanced big data", Fuzzy Sets and Systems, vol. 258, pp. 5–38, January 2015.

[2] Emad A Mohammed, Behrouz H Far, and Christopher Naugler, "Applications of the Map Reduce programming framework to clinical big data analysis: current landscape and future trends", BioData Mining, vol. 7, no.1, 2014.

[3] Sara del Río, Victoria López, José Manuel Benítez, and Francisco Herrera , "On the use of Map Reduce for imbalanced big data using Random Forest", Journal of Information Sciences, vol.285, pp.112–13720, November 2014.

[4] Shamsul Huda, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, and Michael Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis", IEEE Access, Vol. 4, pp. 9145 - 9154, 2017.

[5] Magnus Orn Ulfarsson, Frosti Palsson, Jakob Sigurdsson, and Johannes R. Sveinsson, " Classification of Big Data with Application to Imaging Genetics", Computer Vision and Pattern Recognition, 2016.

[6] Alberto Fernández, Sara del Río, Nitesh V. Chawla, and Francisco Herrera, "An insight into imbalanced Big Data classification: outcomes and challenges", Complex & Intelligent Systems, pp 1–16, 2017.

[7] Dawen Xia, Huaqing Li, Binfeng Wang, Yantao Li, and Zili Zhang, "A MapReduce-Based Nearest Neighbor Approach for Big-Data-Driven Traffic Flow Prediction", IEEE Access, vol. 4, pp. 2920 - 2934, 2016.

[8] Sina Khanmohammadia and Chun-An Choua, "A Gaussian Mixture Model Based Discretization Algorithm for Associative Classification of Medical Data", Expert Systems with Applications, vol. 58, pp. 119–129, 1 October 2016.

[9] Sara del Rio, Victoria Lopez , Jose Manuel Benitez, and Francisco Herrera, "A Map Reduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules", International Journal of Computational Intelligence Systems, vol. 8, no. 3, pp.422-437, 2015.

[10] IBM, What is big data? Bringing big data to the enterprise, [Online; accessed December 2013], http://www-01.ibm.com/software/data/bigdata/, 2012.

[11] P.Zikopoulos, C.Eaton, D.DeRoos, T.Deutsch, G.Lapis, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill, 2011.

[12] S. Madden , "From data bases to big data", IEEE Internet Computing, vol.16, no.3, pp. 4– 6, 2012.

[13] A. Sathi, "Big Data Analytics: Disruptive Technologies for Changing the Game", MCPress, 2012.

[14] Miner, A. Shook, "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems", O'Reilly Media, 2012

[15] Kaplan R.S, Porter M.E, "How to solve the cost crisis in health care", Harv Bus Rev, vol. 89, no.9, pp.46–52, 2011.

[16] Musen M.A, Middleton B, Greenes R.A, "Clinical decision-support systems", Biomedical Informatics, pp. 643–674, 2014.

[17] Devaraj S, Ow TT, Kohli R, "Examining the impact of information technology and patient flow on healthcare performance: A Theory of Swift and Even Flow (TSEF) perspective", Journal of Operations Management, vol. 31, no.4, pp.181–192, May 2013.

[18] Friedman A.B, "Preparing for responsible sharing of clinical trial data", New England Journal of Medicine, vol.370, no. 5, pp.484–484, 2014.

[19] Mazurek M, "Applying NoSQL Databases for Operationalizing Clinical Data Mining Models", International Conference: Beyond Databases, Architectures and Structures, pp 527-536, 2014.

[20] Vijay Mahadeo Mane and D.V. Jadhav, " Holoentropy enabled-decision tree for automatic classification of diabetic retinopathy using retinal fundus images", Biomed. Eng.-Biomed. Tech. 2016.

[21] Hari Singh, Seema Bawa , "A MapReduce-based scalable discovery and indexing of structured big data", Future Generation Computer Systems , vol.73, pp.32-43, August 2017.

[22] Alessio Bechini, Francesco Marcelloni, Armando Segatori, " A MapReduce solution for associative classification of big data", Information Sciences, vol. 332, pp. 33-55, 1 March 2016.

[23] Seema Maitrey, C.K. Jha, " MapReduce: Simplified Data Analysis of Big Data", Procedia Computer Science, vol. 57, pp. 563-571, 2015.

[24] Jin Qian, Ping Lv, Xiaodong Yue, Caihui Liu, Zhengjun Jing, " Hierarchical attribute reduction algorithms for big data using MapReduce", Knowledge-Based Systems, vol. 73, pp.18-31, January 2015.

[25] Cen Chen, Kenli Li, Aijia Ouyang, Keqin Li, " A parallel approximate SS-ELM algorithm based on Map Reduce for large-scale datasets", Journal of Parallel and Distributed Computing, 21 January 2017.

[26] UCI machine learning repository dataset - https://archive.ics.uci.edu/ml/datasets.html.

[27] T. Surekha, Dr. R. Siva Rama Prasad,, " LDT: Log Decision Tree to clinical data classification", Journal of Theoretical and Applied Information Technology, 15 January 2018.

[28] T. Padmapriya, V.Saminadan, "Performance Improvement in long term Evolution-advanced network using multiple imput multiple output technique", Journal of Advanced Research in Dynamical and Control Systems, Vol. 9, Sp-6, pp: 990-1010, 2017.

[29] S.V.Manikanthan and K.srividhya "An Android based secure access control using ARM and cloud computing", Published in: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on 26-27 Feb. 2015,Publisher: IEEE,DOI: 10.1109/ECS.2015.7124833.

[30] Rajesh, M., and J. M. Gnanasekar. "Path observation-based physical routing protocol for wireless ad hoc networks." International Journal of Wireless and Mobile Computing 11.3 (2016): 244-257.