DOI: 10.21917/ijsc.2023.0422

ILLUSTRATING A SCALABLE ARCHITECTURE-POWERED DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Chellammal Surianarayanan[1], Sharmila Rengasamy[2], M. Baby Nirmala[3] and Pethuru Raj Chelliah[4]

[1]*Centre for Distance and Online Education, Bharathidasan University, India*
[2]*Department of Computer Science, Government Arts and Science College, Srirangam, Tiruchirappalli, India*
[3]*Department of Computer Applications, Holy Cross College, India*
[4]*Edge AI Division, Reliance Jio Platforms Ltd., Bangalore, India*

Abstract

Healthcare information systems typically collect, store and manage various kinds of data such as illness details, clinical history, essential body parameters, health insurance plans, and other related data towards enabling data processing and analytics to arrive at better decision making with all the clarity and alacrity. To reduce the mortality rate due to heart diseases, it is essential to predict the presence of disease in its budding stage itself. Manual extraction of the useful knowledge from historical data is practically tedious and time-consuming. Machine learning (ML) algorithms are being used to detect and predict something useful out of both historical and current data. Despite the applicability of machine learning algorithms for prediction, the accuracy of prediction is significantly influenced by features used for prediction. Moreover, to meet the needs of evolving data sizes, suitable technologies for data storage also become essential. Based on these two aspects, a comparative analysis has been performed for feature selection using four filter methods, namely, correlation measure, information gain, gain ratio and relief. Further, a scalable architecture using Hadoop framework has been proposed to enable the machine learning algorithms to handle larger datasets while performing prediction task. The impact of the proposed architecture on the performance of machine learning algorithm has been evaluated with benchmark dataset and found to have improved scalability and accuracy.

Keywords:

Disease Prediction, Hadoop Distributed File System, Machine Learning, Random Forest, Support Vector Machine, Scalable Architecture

1. INTRODUCTION

Heart disease is one of the most significant causes of mortality in the world today and around 17.5 million deaths occur over the world due to heart related diseases [1]-[2]. Machine learning algorithms are predominantly used to detect and predict the presence of the diseases [3]-[6] in individuals at their early stage using the data collected by healthcare information systems. The algorithms are trained using the previously collected historical data so that with the acquired knowledge, they can mimic human brain while classifying the unseen data. Ultimately, they help in decision making and diagnosis of the diseases. The major advantage of machine learning algorithms is that they can be programmed to analyze huge amount of data automatically without human intervention. The time involved in analysis is significantly reduced. In fact, manual extraction of such useful knowledge or decision is infeasible. Machine learning techniques serve as a boon to physicians in relieving their burden in diagnosing the diseases. Despite the availability of many machine learning techniques for disease prediction, the accuracy and performance of prediction are still required to be enhanced as the enhancement will help in saving the human life.

Moreover, machine learning techniques are really evolving with the advancement in hardware and software technologies. As far as prediction is concerned, the traditional relational database systems have limited capability to handle big data due its huge volume, variety, and velocity. The key point here is that the usefulness of big data technologies and tools for prediction of diseases needs to be studied. Keeping these two aspects in mind, a big data-based architecture is proposed for disease prediction using supervised machine learning techniques. The primary motivation of the work is to utilize effectively, the facility provided by big data platforms to store and process massive amounts of data. The proposed architecture has been designed using Hadoop Distributed File System (HDFS). The unique feature of HDFS is that it can be deployed on low priced commodity hardware. HDFS is highly fault tolerant and also, it is suitable for large dataset and provides high throughput access to application data.

The objective of this work is to analyze the performance of machine learning algorithms in predicting the diseases in the proposed architecture with respect to accuracy, scalability and computation time. Another objective is to perform a comparative analysis on feature selection using four filter methods namely correlation measure, information gain, gain ratio and relief. Based on these aspects a two-phase methodology has been proposed for prediction of heart disease using machine learning algorithms in a scalable architecture. The major contributions include:

- Recommendation of relevant features and algorithms after performing comparative analysis of mention methods using three supervised algorithms, Random Forest (RF), Support Vector Machine (SVM) and J48 classifiers.
- Construction of Hadoop and Spark based architecture for storage and processing of large-scale data in three different configurations namely single node, standalone cluster and distributed cluster.
- Evaluation of performance of prediction algorithms in the proposed architecture and recommendation of suitable algorithm for large scale data

In contrast to the existing similar research works, the present work is unique in analyzing the performance of machine learning algorithms in a scalable architecture. More specifically, through this extensive study, the Random Forest algorithm is found to give the best accuracy and less computation time for both small- and large-scale datasets. The presented scalable architecture simplifies the validation of different machine learning models with larger datasets.

2. LITERATURE SURVEY

The research works that have theme like the present work are reviewed. In [7], the authors proposed an integrated approach consisting of Apache Hive and Tableau (analysis tool) extract useful information from Electronic Health Record (EHR) dataset collected from Open Data Commons Open Database License (ODbL) and stored in Hadoop Distributed File System (HDFS). In [8], the author's presented a big data-based approach to detect the presence of heart diseases. The data for analysis is loaded into HDFS and queried using Hive. In addition, impala, an open-source parallel processing Structured Query Language (SQL) engine is used for prediction. In [9], Convolutional Neural Network based multimodal disease prediction algorithm is used for a disease prediction from large volume of data collected from a real hospital. In [10], the authors discussed an early heart disease detection using data mining techniques with Hadoop and MapReduce. The accuracy of MapReduce algorithm was found to yield better accuracy than K-Means algorithm.

In [11], a model has been discussed to predict heart disease using Artificial Neural Network (ANN) and MapReduce algorithms. In [12], an approach has been proposed for predicting heart failure by using multi-structure dataset integrated from various resources which was managed using HDFS. In [13], cluster analysis has been performed with MapReduce programming models with specific focus on privacy. In [14], the authors proposed a real-time monitoring and scalable system for early detection of heart disease using Spark and Cassandra frameworks. In [15], the authors designed probabilistic classification using Bayes theorem in MapReduce programming paradigm. In [16], MapReduce based disease prediction for various disease occurrences using decision tree has been suggested to enhance efficiency. In [17], MapReduce based centralized patient monitoring system has been presented to analyze vital features such as respiratory rate interval, QRS interval and QT interval to decide whether the monitored features are normal or abnormal. In [18], prediction of the Coronary Artery has been performed using big data platforms. Further, big data-based platforms can extract valuable insights out of massive, complex, interconnected unstructured data [12]. Moreover, real time monitoring, and analytics are facilitated by distributed frameworks such as Apache Spark which provides an inbuilt machine learning library [19].

3. PROPOSED METHODOLOGY

The proposed methodology consists of two phases, phase-1 deals with selection of relevant features and prediction algorithms and phase-2 deals with the construction big data-based architecture and performing prediction task using selected features and algorithms, resulted from phase-1, in the constructed architecture.

3.1 PHASE 1 - SELECTION OF RELEVANT FEATURES AND PREDICTION ALGORITHMS

The accuracy of prediction is affected by redundant, irrelevant, and not useful features [20]. In this work, filter methods are used as

- They are independent of prediction algorithms.
- They are fast and computationally cheap.
- They easily scale to high-dimensional data.
- They have potential for good generalization.

Features are selected based on statistical measures. In the proposed method four statistical measures, namely, correlation measure, information gain, gain ratio and relief and three supervised algorithms, namely, RF, SVM and Relief are used. At first, weight (or rank) of features in the dataset have been found out using the above measures. Secondly, the features are added one by one according to their weight and each time the accuracy of prediction of different algorithms have been found out.

Phase 1 includes different tasks namely identification of benchmark dataset, selection of splitting criteria, selection of prediction algorithms, performing experiments. The obtained results are inter-compared to find the out the relevant features and more suitable algorithms. The results of phase 1 are given to phase 2. The block diagram of phase 1 is shown in Fig.1.

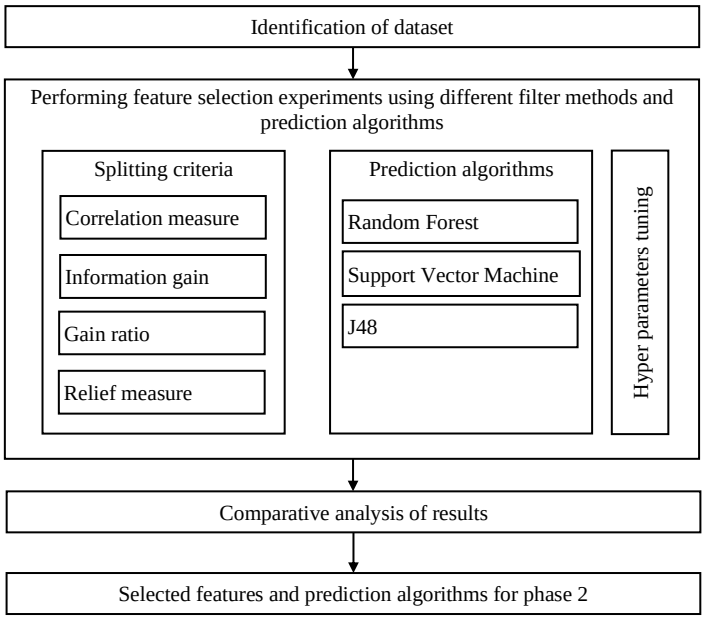

Fig.1. Block diagram of phase 1 of the proposed method

3.1.1 Dataset:

Benchmark dataset from University of California Irvine (UCI) machine learning repository [21] has been chosen. In this work heart disease dataset has been collected from Cleveland Clinic Foundation, consisting of 303 records have been chosen. There were 6 records having missing values and those records have been eliminated. The dataset contains 13 features and one targeted class (called as 'num'). Details about the features are given in Table.1.

3.1.2 Methods for Feature Selection:

As mentioned earlier, four statistical measures, namely, correlation measure, information gain and gain ratio and relief are used to find the relationship between the features and the target variable. They are computed as follows.

Correlation measure - Pearson's Correlation is used as measure to find out the linear dependence between two variable using the formula given in Eq.(1)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i \cdot \bar{x})(y_i \cdot \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i \cdot \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i \cdot \bar{y})^2}} \quad (1)$$

In Eq.(1), r_{xy} denotes the correlation co-efficient between x and y. x_i denotes the i^{th} sample of x, y_i denotes the i^{th} sample y, $\bar{x}$ denotes the mean of x and $\bar{y}$ denotes the mean of y. Also, n denotes the number of samples.

Information Gain: It refers to the expected reduction in entropy caused by partitioning the data according to a particular attribute. It is computed using Eq.(2)

$$gain(T,a) = entropy(T) - \sum_{i=1}^{|a|}\frac{|a_i|}{|T|}entropy(a_i) \quad (2)$$

In (2), $gain(T,a)$ represents the information gain provided by the feature a about the target class, T, $entropy(T)$ denotes the entropy of the target class label, and $\sum_{i=1}^{|a|}\frac{|a_i|}{|T|}entropy(a_i)$ denote the entropy of the dataset, given the variable, 'a'. The value of $entropy(T)$ is computed using Eq.(3)

$$entropy(T) = -\sum_{n=1}^{N} p_i \log_2 p_i \quad (3)$$

In Eq.(3), N is the number of classes, and P_i is the frequency of class i in the same dataset.

Gain ratio: The value of gain ratio is computed using (4)

$$gain\_ratio(T,a) = \frac{entropy(T) - \sum_{i=1}^{|a|}\frac{|a_i|}{|T|}entropy(a_i)}{entropy(T)} \quad (4)$$

Relief: Relief is an instance-based feature selection method which evaluates a feature by how well its value distinguishes samples that are from different groups but are like each other. For each feature X, Relief-F selects a random sample and k of its nearest neighbors from the same class and each of different classes. Then X is scored as the sum of weighted differences in different classes and the same class. If X is differentially expressed, it will show greater differences for samples from different classes, thus it will receive higher score (or vice versa).

3.1.3 Methods for Feature Selection:

Three prediction algorithms, RF, SVM and J48 decision tree are chosen for prediction of heart diseases. Random Forest algorithm is based on the concept of ensemble learning [22]. It combines multiple decision trees to predict the class of the dataset. It takes less training time than other algorithms and it runs effectively for large dataset, and it also maintains high accuracy when large portion of dataset is missing [23]. In RF randomly selected features set is used to split each node [24]. SVM is used for both classification and regression even in complex domains. The main aim of SVM is to find the best hyper plane that separates all data points of one class from the other class by creating a margin between two classes [25]. Margin refers to the distance between hyper plane and the nearest support vector. The distance of margin should be as large as possible. If the data points are not linearly separable, then kernel functions are used to map the non-linearity in input to linear data in a high dimensional space with the help of mathematical functions. As far as heart disease prediction is concerned, polynomial, Radial Basis Function (RBF), sigmoid and linear are extensively used [26]. J48 Choice tree is the usage of calculation ID3 (Iterative Dichotomise variant 3) and it is used for heart disease prediction as mentioned in [27].

3.1.4 Experiments:

Experiments have been performed using Weka tool on a Pentium dual core processor having 4 GB RAM. At first the weights of features have been obtained using different measures and given in Table.2. Secondly, prediction of heart diseases using different algorithms by including features one by one according to the weights obtained using correlation measure has been performed. The accuracy of prediction of different classifiers obtained with correlation measure is given in Table.3. Similarly, prediction of heart diseases using different algorithms by including features one by one according to the weights obtained using information gain, gain ratio and relief measure has been performed. The accuracy of prediction of different classifiers obtained with these measures are given in Table.4, Table.5 and Table.6, respectively.

Another key point to be noted is that the algorithms have been tuned for their hyperparameters as given below. The accuracy of SVM is found to be the best with sigmoid kernel function when compared with other kernel functions, linear, polynomial and radial basis. Similarly, for random forest, the number of decision trees has been chosen as 25. The depth of tress is fixed as 4. Also, the split ratio of training to testing is 70:30.

The following inferences are drawn from Table.3, Table.4, Table.5 and Table.6.

- The accuracy of RF and SVM are found to be higher than J48.
- The algorithms are found to give optimal accuracy for the same set of attributes, namely, thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg for both correlation measure and information gain.
- The accuracy of the algorithms is found to be higher with the features selected by correlation measure and information gain than the other two measures.
- As both the statistical measures, correlation and information gain identifies the same set of features and as the algorithms are found to give the optimal accuracy for the above features, these features (thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg) are recommended as suitable features for phase 2 of the work. The three attributes tresbps, chol and fbs are found to be redundant and irrelevant.
- As the accuracy of J48 is reasonably lower that the other two algorithms, it is not considered for phase 2.

Table.1. Description about features

S. No	Attribute	Value	Description
1	Age	29 – 62	Age in years
2	Sex	0 – male 1- female	Gender
3	Cp	1-typical angina; 2-atypical angina; 3-non-anginal pain; 4-asymptomatic	Chest pain type
4	Trestbps	Numeric value (140mm/Hg)	Resting blood pressure in mm/Hg. Blood pressure should be less than 120/80mm/Hg.
5	Chol	Numeric value (289mg/dl)	Serum cholesterol in mg/dl.
6	Fbs	1-true, 0-false	Fasting Blood sugar>120mg/dl
7	Restecg	0-normal, 1-having ST-T, 2-hypertrophy	Resting electrocardiographic Results should lie between 0 and 2.
8	Thalach	140,173	Maximum heart rate
9	Exang	1-yes, 0-no	Exercise induced angina
10	Oldpeak	Numeric value	ST depression induced by exercise
11	Slope	1-upsloping, 2-flat, 3-downsloping	The slope of the peak exercise ST segment
12	Ca	0-3 vessels	Number of major vessels colored by fluoroscopy.
13	Thal	3-normal, 6-fixed defect, 7-reversable defect	Thalassemia (It is a type of blood disorder which reduces the ability of body in producing hemoglobin.
14	Num	0: < 50% diameter narrowing 1: > 50% diameter narrowing	diagnosis of heart disease (angiographic disease status)

Table.2. Weights of features according to different measures

Correlation measure		Information gain		Gain ratio		Relief measure	
Attribute	**Value**	**Attribute**	**value**	**Attribute**	**value**	**Attribute**	**Value**
Thal	0.4862	Thal	0.211628	Ca	0.174093	Cp	0.17195
Ca	0.4608	Cp	0.204599	Thal	0.169789	Thal	0.12783
Exang	0.4368	Ca	0.17025	Exang	0.156009	Sex	0.11551
Oldpeak	0.4307	Oldpeak	0.15845	Thalach	0.132156	Ca	0.0923
Thalach	0.4217	Exang	0.14221	Cp	0.117624	slope	0.07657
Cp	0.3817	Thalach	0.129652	Oldpeak	0.105318	exang	0.06568
Slope	0.3564	Slope	0.116834	Slope	0.090289	restecg	0.0637
Sex	0.2809	Age	0.060167	Sex	0.065643	oldpeak	0.02355
Age	0.2254	Sex	0.059138	Age	0.060273	Fbs	0.02079
Restecg	0.1664	Restecg	0.024075	Restecg	0.022129	thalach	0.02006
Trestpbs	0.1449	Fbs	0.000566	Fbs	0.000934	Age	0.01484
Chol	0.0852	Trestbps	0	Chol	0	trestbps	0.01446
Fbs	0.028	Chol	0	Trestbps	0	Chol	-0.001

Table.3. Accuracy of different classifiers with correlation measure

Attributes	**Accuracy (in %)**		
	RF	**SVM**	**J48**
Thal	76.4310	76.5677	76.5677
Thal, ca	78.1145	75.5776	77.8878
Thal, ca, exang	83.5017	78.5479	82.1782
Thal, ca, exang, oldpeak	80.4714	80.8581	77.2277
Thal, ca, exang, oldpeak, thalach	81.8182	84.1584	76.5677
Thal, ca, exang, oldpeak, thalach, cp	82.4916	83.8284	76.5677
Thal, ca, exang, oldpeak, thalach, cp, slope	82.8283	83.8284	76.8977
Thal, ca, exang, oldpeak, thalach, cp, slope, sex	82.1549	83.4983	77.5578
Thal, ca, exang, oldpeak, thalach, cp, slope, sex, age	83.5017	83.4983	75.9076
Thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg	**84.5118**	**84.8185**	**77.5578**
Thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps	83.5017	84.8185	77.2277
Thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol	81.8182	84.4884	77.5578
Thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol, fbs	83.5017	84.1584	77.5578

Table.4. Accuracy of different classifiers with information gain

Attributes	**Accuracy (in %)**		
	RF	**SVM**	**J48**
Thal	76.5677	76.5677	76.5677
Thal, cp	73.2673	72.2772	72.9373
Thal, cp, ca	82.8383	77.5578	81.5182
Thal, cp, ca, oldpeak	77.8878	82.8383	81.1881
Thal, cp, ca, oldpeak, exang	78.8779	83.1683	79.868
Thal, cp, ca, oldpeak, exang, thalach	80.5281	83.8284	76.5677
Thal, cp, ca, oldpeak, exang, thalach, slope	79.2079	83.8284	76.8977
Thal, cp, ca, oldpeak, exang, thalach, slope, age	82.5083	84.1584	75.2475
Thal, cp, ca, oldpeak, exang, thalach, slope, age, sex	83.8482	83.4983	75.9076
Thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg	**84.5082**	**84.8185**	**77.5578**
Thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg, fbs	83.5182	83.4983	76.8977
Thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg, fbs, trestbps	82.5083	83.8284	77.2277
Thal, cp, ca, oldpeak, exang, thalach, slope, age, sex, restecg, fbs, trestbps, chol	81.1683	84.1584	77.5578

Table.5. Accuracy of different classifiers with gain ratio

Attributes	**Accuracy (in %)**		
	RF	**SVM**	**J48**
Ca	74.5875	68.6469	74.5875
Ca, thal	78.2178	75.5776	77.8878
Ca, thal, exang	78.8284	78.5479	82.1782
Ca, thal, exang, thalach	79.2475	82.5083	77.5578
Ca, thal, exang, thalach, cp	79.9076	83.8284	79.538
Ca, thal, exang, thalach, cp, oldpeak	80.2281	83.8284	76.5677
Ca, thal, exang, thalach, cp, oldpeak, slope	80.3079	83.8284	76.8977
Ca, thal, exang, thalach, cp, oldpeak, slope, sex	81.5182	83.4983	76.9578
Ca, thal, exang, thalach, cp, oldpeak, slope, sex, age	81.8482	83.4983	77.2076

Ca, thal, exang, thalach, cp, oldpeak, slope, sex, age, restecg	**81.8482**	**83.8185**	**75.5578**
Ca, thal, exang, thalach, cp, oldpeak, slope, sex, age, restecg, fbs	81.5182	83.4983	76.8977
Ca, thal, exang, thalach, cp, oldpeak, slope, sex, age, restecg, fbs, chol	81.1782	82.1584	76.2277
Ca, thal, exang, thalach, cp, oldpeak, slope, sex, age, restecg, fbs, chol, trestbps	81.1683	81.1584	77.5578

Table.6. Accuracy of different classifiers with relief measure

Attributes	**Accuracy (in %)**		
	RF	**SVM**	**J48**
Cp	75.9076	75.9076	75.9076
Cp, thal	73.2673	72.2772	72.9373
Cp, thal, sex	75.9076	72.2772	74.9175
Cp, thal, sex, ca	80.8581	83.1683	79.868
Cp, thal, sex, ca, slope	80.5281	84.1584	82.5083
Cp, thal, sex, ca, slope, exang	81.1881	84.4884	78.8779
Cp, thal, sex, ca, slope, exang, restecg	81.8482	83.4983	78.8779
Cp, thal, sex, ca, slope, exang, restecg, oldpeak	81.8482	84.8185	79.2079
Cp, thal, sex, ca, slope, exang, restecg, oldpeak, fbs	81.1881	84.1584	79.2079
Cp, thal, sex, ca, slope, exang, restecg, oldpeak, fbs, Thalach	**82.5083**	**83.8284**	**78.5479**
Cp, thal, sex, ca, slope, exang, restecg, oldpeak, fbs, Thalach, age	81.5182	83.4983	76.8977
Cp, thal, sex, ca, slope, exang, restecg, oldpeak, fbs, Thalach, age, trestbps	82.5083	83.1284	77.2277
Cp, thal, sex, ca, slope, exang, restecg, oldpeak, fbs, thalach, age, trestbps, chol	83.1683	84.1584	77.5578

3.2 PHASE 2 – CONSTRUCTION OF SCALABLE ARCHITECTURE AND PERFORMING EXPERIMENTS

3.2.1 Construction of Scalable Architecture:

Scalability is an important aspect to be considered in any mining task. In modern era, data is being generated from different sources continuously and it becomes essential to study the performance of prediction algorithms in large scale datasets. Conventional database systems have shortcomings in handling the large size datasets.

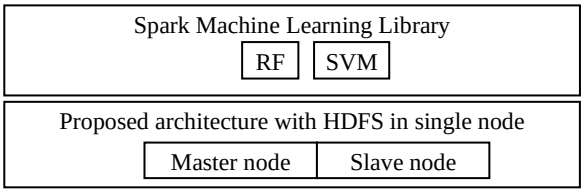

Fig.2. Block diagram of single node configuration

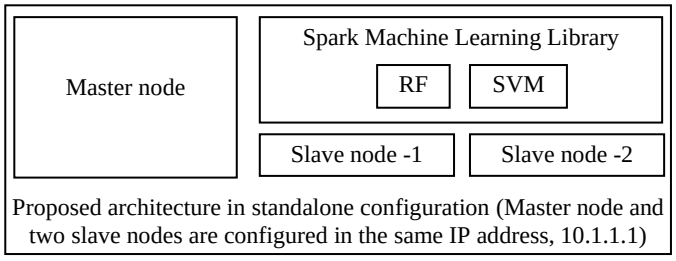

Fig.3. Block diagram of standalone cluster configuration

In this work, Hadoop Distributed File System (HDFS) has been used for storage. In HDFS, there are three important nodes, master node, secondary master node and slave node. Master node stores the metadata of the data (in two different files, editlogs and fsimage), manages the data storage in different slave (or data) nodes. The main function of secondary master node is to checkpoint in HDFS. The slave or data nodes store the data. HDFS stores data in the form of block of 64MB or 128MB. Data is stored as simple files with append option. There is no update of data. Thus, it is very opted for performing mining task. Machine learning libraries provided by Apache Spark distributed framework has been used in this work. The Spark is more efficient when compared to Mapreduce due to the unique feature of Spark that is has large in-memory capacity. So, the number of data fetches required for Spark is very low. It is proposed to perform the prediction task in three different configurations, namely single node configuration, standalone node configuration and distributed node configuration. The configurations are described below. The block diagram of the proposed architecture in the above configurations are given in Fig.2, Fig.3 and Fig.4.

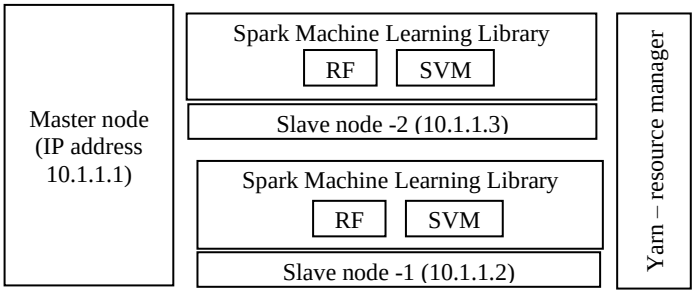

Fig.4. Block diagram of distributed cluster configuration

- *Single node configuration*: This configuration refers to the process of performing disease prediction in a single node having HDFS file system. Only one data node is involved, and it is configured in the same node where master node resides.
- *Standalone Cluster*: In standalone cluster configuration, prediction process is carried out in a cluster consisting of one master and more than one slave node. In this work, two slave nodes are used. In this configuration, the slave nodes and master node are configured in the same IP address.
- *Distributed Cluster*: In distributed cluster configuration, prediction process is carried out in a cluster of one master and more than one slaves which are configured in different IP addresses in a distributed manner. In this work, one master and two slaves are used, and they are configured in different IP address.

3.2.2 Evaluation of Prediction in the Proposed Architecture:

The objective of the experimentation is to evaluate the performance of the prediction with large size data. The data is stored in HDFS. Datasets having larger number of records (of 1, 40,000 records and of the order of 10MB to 100MB) have been generated from the original dataset containing 297 records as the base by randomly varying the record level data. Variation in accuracy and computation time have been analyzed by increasing the number of records from 297 to 1, 00, 000(1 lakh) as well as of the order of Megabytes from 10 MB to 100MB. Default block size of 64MB has been used during experimentation.

Three experiments have been conducted. In the first experiment, the performance of classifiers has been analyzed with respect to two file systems, namely, normal FAT32 file system and HDFS (i.e. single node configuration). In the second and third experiments, the performance of the classifiers have been analyzed in standalone cluster and distributed cluster configurations. The performance has been evaluated using accuracy and computation time.

Results obtained in single node configuration - The accuracy and computation time for classifiers by varying the number of records varied from 297 to 1, 40,000 are given in Table.7. From Table.7 the following key points have been inferred.

- It is found that the accuracy obtained using RF is higher when compared to that of SVM both in conventional and distributed file system. In addition, conventional file system could not produce results when the number of records exceeds 1 lakh whereas HDFS is found to perform well even when the number of records exceeds 1 lakh.
- The accuracy of RF is found to be higher than that of SVM both in normal and HDFS.
- The computation time obtained using RF is lower when compared to that of SVM both in normal and HDFS.
- More importantly the scalability is enhanced in HDFS where the normal file system is found to be not able to produce response when the record size exceeds 1 lakh.

Results obtained in standalone cluster configuration – The accuracy and computation time of the classifiers in standalone cluster configuration by varying the number of records from 297 to 1,40,000 are given in Table.8. From Table.8, the accuracy of RF is found to be higher than that of SVM. The computation time of distributed cluster is almost comparable with that of standalone cluster in spite of the inter-node communication in distributed cluster whereas in standalone it is only inter-process.

More important aspect to be noted is the stability in accuracy and computation time of RF. The variations in accuracy and computation time are very minimum for the wide range record size from 297 to 140000. This implies that RF is very efficient in handling large size data. It is basically arising from the principle of working of RF. RF constructs various decision tree by using sampling data and it does not take the entire data into consideration. It selects the subsets of data at random also with randomly selected feature sets. Also, the number of trees can be tuned as a hyperparameter. Since it is an ensemble of trees its accuracy is also high.

Ultimately another experiment has been conducted with RF alone by varying the input size of data of 10MB to 100MB. The variations in accuracy and computation time of RF are given in Table.9.

From Table.9, Random Forest algorithm produces high accuracy with increase in data size from 10MB to 100 MB. More importantly the accuracy and computation time is found to be very stable. In addition, the computation time of RF in distributed cluster is also almost comparable to that of standalone. A few seconds of difference in time in distributed cluster is due to the communications performed by YARN. Ultimately, from experimentation, RF is found to give better and stable performance for prediction of diseases in large scale datasets.

Table.7. Accuracy and computation time of classifiers in normal file system and HDFS

Number of Records	Accuracy (%)				Computation time (seconds)			
	Normal File System		HDFS		Normal File System		HDFS	
	RF	SVM	RF	SVM	RF	SVM	RF	SVM
297	0.81	0.8666	0.7888	0.8666	0.0580	0.0260	7.8861	7.7491
5049	0.8884	0.7570	0.8739	0.7570	0.0840	0.1720	8.2968	8.4792
10098	0.8818	0.7663	0.8927	0.7663	0.1139	0.6519	8.8410	9.8212
20196	0.8978	0.7562	0.8942	0.7562	0.1679	2.5829	9.3899	15.1608
40392	0.8988	0.7601	0.9017	0.7601	0.2579	10.2480	11.0194	35.1571
60000	0.9000	0.763	0.8923	0.763	0.3689	21.1879	22.1534	70.3288
80784	0.8999	0.7569	0.8963	0.7569	0.5130	35.9419	13.6792	111.5218

100000	0.8921	0.7672	0.8902	0.7672	0.5770	61.5230	14.4925	186.0040
120000	No response		0.8990	0.7699	No response		15.7865	260.0890
140000	No response		0.9061	0.7754	No response		16.5432	350.7651

Table.8. Accuracy and computation time of classifiers in standalone cluster and distributed cluster

No of Records	Accuracy (%)				Computation time (seconds)			
	Standalone cluster		Distributed cluster		Standalone cluster		Distributed cluster	
	RF	SVM	RF	SVM	RF	SVM	RF	SVM
297	0.8333	0.8666	0.8	0.8666	31.2816	17.7179	13.7392	17.2427
5049	0.8805	0.7570	0.8838	0.7570	23.5762	12.2468	26.7426	16.3527
10098	0.8924	0.7663	0.8897	0.7663	35.6660	13.5834	45.1570	17.4452
20196	0.8999	0.7562	0.8958	0.7562	24.9411	18.7990	27.5334	33.3564
40392	0.8955	0.7601	0.9018	0.7601	33.2358	37.9331	43.5633	46.0661
60000	0.9128	0.763	0.9098	0.763	26.1018	134.9976	23.9547	78.7345
80784	0.8964	0.7569	0.8930	0.7569	26.1583	142.577	34.2709	121.8103
100000	0.8973	0.7672	0.8852	0.7672	24.9004	203.7866	41.9106	209.2389
120000	0.9017	0.7712	0.8954	0.7692	25.1201	264.8645	40.8967	269.4329
140000	0.9062	0.7732	0.8992	0.7708	25.1243	299.9345	39.1023	287.7107

Table.9 Accuracy and Computation time of RF

	Accuracy (%)		Computation time (seconds)	
	Stand-alone cluster	Distributed cluster	Stand-alone cluster	Distributed cluster
10MB	0.8809	0.8980	72.2884	76.6543
20MB	0.9020	0.9020	98.5748	102.0123
30MB	0.9051	0.9051	99.1594	103.1320
40MB	0.9123	0.9123	100.8261	106.5642
50MB	0.9124	0.9124	99.6092	108.9231
60MB	0.9167	0.9145	101.4591	110.1762
70MB	0.9179	0.9167	102.5798	112.8342
80MB	0.9196	0.9187	100.6432	116.9815
90MB	0.9276	0.9193	102.9810	118.7645
100MB	0.9298	0.9197	101.9765	120.2456

4. CONCLUSION

Nowadays data is getting generated in huge volume in healthcare domain. Healthcare information systems archive data related to personal information such as EHR, data related to treatment, data archived from various medical tests, clinical reports, and other medical data from social network, smart phones, etc. The main point is that the data size is keeping on growing. So, a scalable architecture and efficient machine learning algorithms are necessary to perform prediction of diseases. Also, the accuracy of any prediction algorithm is influenced by the features involved. A comparative analysis has been perform using four filter methods and selection of right features is ensured. A scalable architecture using HDFS file system and Spark distributed framework has been proposed. The architecture has been established in three different configurations namely single node configuration, standalone configuration, and distributed cluster configuration. Elaborate experimentation has been performed by varying the data sizes to analyze the performance of machine learning algorithms for their performance in terms of accuracy and computation time. It has been found that the algorithms do not produce any results in normal file system when the number of records exceeds 1 lakh. This proves the need of scalable architecture. Further the performance of SVM and RF are analyzed in standalone and distributed cluster configurations. The accuracy and computation time of RF is found to be higher and sTable.when compared to that of SVM.

REFERENCES

[1] Senthilkumar Mohan and Gautam Srivastava, “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”, *IEEE Access*, Vol. 7, pp. 81542-81553, 2019.

[2] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, "Heart Disease Diagnosis and Prediction using Machine Learning and Data Mining Techniques: A Review", *Advances in Computational Sciences and Technology*, Vol. 10, pp. 2137-2159, 2017.

[3] Apurb Rajdhan, Milan sai, Avi Agarwal, Dundigalla Ravi and Poonam Ghuli, "Heart Disease Prediction using Machine Learning", *International Journal of Engineering Research and Technology*, Vol. 9, No. 4, pp. 659-662, 2020.

[4] N. Arunpradeep and G. Niranjana, "Different Machine Learning Models Based Heart Disease Prediction", *International Journal of Recent Technology and Engineering*, Vol. 8, No. 6, pp. 544-548, 2020.

[5] Raparthi Yaswanth and Y.M. Riyazuddin, "Heart Disease Prediction using Machine learning Techniques", *International Journal of Innovative Technology and Exploring*, Vol. 9, No. 5, pp. 1456-1460, 2020.

[6] Rajatdeep Kaur and Kamaljit Kaur, "Cardiovascular Disease Recognition through Machine Learning Algorithms", *International Journal of Engineering and Advanced Technology*, Vol. 9, No. 4, pp. 2109-2115, 2020.

[7] Zhenlin Kan, Xinru Cheng, Seung Hyun Kim and Yuting Jin, "Apache Hive-Based Big Data Analysis of Healthcare", *International Journal of Pure and Applied Mathematics*, Vol. 119, No. 8, pp. 237-259, 2018.

[8] Niha Beera, Nysha Chaparala and Jaya Lakshmi Gundabathina, "Data Analysis of Heart Disease Dataset using Hadoop and Impala with MySQL", *International Journal of Applied Engineering Research*, Vol. 13, No. 7, pp. 5311-5315, 2018.

[9] Shraddha Subhash Shirsath and Pro. Shubhangi Patil, "Disease Prediction using Machine Learning over Big Data", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 7, No. 6, pp. 6752-6757, 2018.

[10] S. Bagavathy, V. Gomathy, S. Sheeba Rani, and Monica Murugesan, "Early Heart Disease Detection using Data Mining Techniques with Hadoop Map Reduce", *International Journal of Pure and Applied Mathematics*, Vol. 119, No. 12, pp. 1915-1920, 2018.

[11] T. Nagamani, S. Logeswari and B. Gomathy, "Heart Disease Prediction using Data Mining with Mapreduce Algorithm", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8, No. 3, pp. 1-13, 2019.

[12] Heba F. Rammal and Ahmed Z. Emam, "Heart Failure Prediction Models using Big Data Techniques", *International Journal of advanced Computer Science and Applications*, Vol. 9, No. 5, pp. 363-371, 2018.

[13] S. Yamini and K.P. Rama Prabha, "A Data Mining with Big Data Disease Prediction", *International Research Journal of Engineering and Technology*, Vol. 5, No. 4, pp. 829-832, 2018.

[14] Abderrahmane ED Daoudy and Khalil Maalmi, "Real-Time Machine Learning for Early Detection of Heart Disease using Big Data Approach", *Proceedings of IEEE International Conference on Wireless Technologies, Embedded and Intelligent Systems*, pp. 1-6, 2019.

[15] R. Venkatesh, C. Balasubramanian and M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine Learning Technique", *Journal of Medical Systems*, Vol. 78, pp. 1-14, 2019.

[16] S. Vinitha and S. Sajini, "Disease Prediction using Machine Learning over Big Data", *Computer Science and Engineering: An International Journal*, Vol. 8, No. 1, pp. 1-8, 2018.

[17] G. Vaishali and V. Kalaivani, "Big Data Analysis for Heart Disease Detection System using Map Reduce Technique", *Proceedings of International Conference on Computing Technologies and Intelligent Data Engineering*, pp. 1-6, 2016.

[18] Prema Jain and Amandeep Kaur, "Big Data Analysis for Prediction of Coronary Artery Disease", *Proceedings of International Conference on Computing Sciences*, pp. 188-193, 2018.

[19] Cheryl Ann Alexander and Lidong Wang, "Big Data Analytics in Heart Attack Prediction", *Journal of Nursing and Care*, Vol. 6, No. 2, pp. 1-9, 2017.

[20] Mohmmed Abdulrazzaq Thanoon, Mohammad J.M. Zedan and Abdulhameed N. Hameed, "Feature Selection Based on Wrapper and Information Gain", *Proceedings of International Conference for Science and Technology*, pp. 1-6, 2019.

[21] Heart Disease Data Set, Available at http.//archive.ics.uci.edu/ml/datasets/Heart+Disease, Accessed at 2022.

[22] Priya R. Patil and S.A. Kinariwala, "Automated Diagnosis of Heart Disease using Random Forest Algorithm", *International Journal of Advance Research, Ideas and Innovations in Technology*, Vol. 3, No. 2, pp. 579-589, 2017.

[23] C. Beulah Christalin Latha and S Carolin Jeeva, "Improving the Accuracy of Prediction of Heart Disease Risk based on Ensemble Classification Techniques", *Informatics in Medicine Unlocked*, Vol. 16, No. 1, pp. 1-14, 2019.

[24] P. Nancy, B. Swaminathan, K. Navina, B. Nandhini and P. Lokesh, "Tuned Random Forest Algorithm for Improved Prediction of Cardiovascular Disease", *International Journal of Recent Technology and Engineering*, Vol. 9, No. 1, pp. 1355-1360, 2020.

[25] K.S. Shalet and V.J. Sarath Kumar, "Diagnosis of Heart Disease using Decision Tree and SVM classifier", *International Journal of Applied Engineering Research*, Vol. 10, No. 68, pp. 598-602, 2015.

[26] Deepika Kancherla, Jyostna Devi Bodapati and N. Veeranjaneyulu, "Effect of Different Kernels on the Performance of an SVM Based Classification", *International Journal of Recent Technology and Engineering*, Vol. 7, No. 4, pp. 1-6, 2019.

[27] K.M. Almustafa, "Prediction of Heart Disease and Classifiers' Sensitivity Analysis", *BMC Bioinformatics*, Vol. 21, pp. 278-289, 2020.