Real-time machine learning for early detection of heart disease using big data approach

Abderrahmane Ed-daoudy*
LTTI, ESTF, Université Sidi Mohamed Ben Abdellah, Route d'Imouzzer, BP 2427, Fès 30000, Morocco
a.eddaoudy@gmail.com

Khalil Maalmi
LTTI, ESTF, Université Sidi Mohamed Ben Abdellah, Route d'Imouzzer, BP 2427, Fès 30000, Morocco
k_maalmi@yahoo.com

***Abstract*— Over the last few decades, heart disease is the most common cause of global death. So early detection of heart disease and continuous monitoring can reduce the mortality rate. The exponential growth of data from different sources such as wearable sensor devices used in Internet of Things health monitoring, streaming system and others have been generating an enormous amount of data on a continuous basis. The combination of streaming big data analytics and machine learning is a breakthrough technology that can have a significant impact in healthcare field especially early detection of heart disease. This technology can be more powerful and less expensive. To overcome this issue, this paper propose a real-time heart disease prediction system based on apache Spark which stand as a strong large scale distributed computing platform that can be used successfully for streaming data event against machine learning through in-memory computations. The system consists of two main sub parts, namely streaming processing and data storage and visualization. The first uses Spark MLlib with Spark streaming and applies classification model on data events to predict heart disease. The seconds uses Apache Cassandra for storing the large volume of generated data.**

Keywords*— real-time, distributed machine learning, big data, spark, heart disease, MLlib

I. INTRODUCTION

The heart is a muscle and its role is to pump the blood throughout the entire body. It constitutes the principal element of the body. Heart disease is one of the leading health risk facing men today. According to the World Health Organization (WHO), stroke and heart attack are the most common cause (80%) of global death [1]. Hence, the availability of data and data mining techniques especially machine learning and early detection of heart disease can help patients react in advance to a probable disease. In healthcare field, due the huge amount of data (big data) generated from multiple areas, multiple sources such as streaming machines, advanced healthcare systems, high throughput instruments, sensor networks, internet of things, mobile application, data collection and processing is becoming very common these days.

Hadoop MapReduce [2][3] with the next generation processing engine for big data, Apache Spark [4] are the most used framework in dealing with big data. One of the primary drawbacks of Hadoop MapReduce is only supports batch processing, it is not suitable for real-time stream processing and in-memory computation.

Apache Spark [4] extends the MapReduce model to support more sophisticated computations. Spark offers a concept named Resilient Distributed Datasets (RDDs) [5] which is the heart of Spark designed to support in-memory data storage and distributed computing. The Spark project stack currently is comprised of Spark Core and four libraries namely MLlib for machine learning and Spark streaming for stream data processing. The algorithms in this library are optimized to run over a distributed dataset, which is more suitable for real-time prediction.

The proposed continuous heart disease monitoring system is arranged as follows: Section 2 reviews the recent works, in this field. The proposed system, result and discussion are described in Section 3 and 4 respectively. Section 5 presents the future work and concludes the paper.

II. RELATED WORK

Nowadays big data analytics especially healthcare analytics has become an important issue for a large number of researches. Recently, many researchers are using machine learning in healthcare.

In [6] an experiment was performed for the prediction of heart attacks and comparison to find the best method of prediction. In [7] a cloud based K-means clustering running as a MapReduce job has been proposed which use healthcare data on cloud for clustering. Authors in [8] have proposed a modern model and system architecture to handle large amounts of information. A web enabled distributed and electronic health record personal health record management framework is proposed using Hadoop and HBase [9].

Usage of convolutional neural network based multimodal disease risk prediction algorithm for disease prediction by machine learning over big data from healthcare communities is performed in [10]. A Hadoop based intelligent care system is proposed in [11] that illustrates internet of things based big data contextual sharing across all devices in a health system. A model for real time analysis of medical big data is proposed in [12]. The approach is exemplified through Spark Streaming and apache Kafka using the processing of healthcare big data Stream. On the other hand stream computing over big data is performed in some papers. For example a prediction approach is proposed in [13], the proposed solution is based on big data processing engine and MLlib. The data is received and filtered from twitter, applying machine learning and send appropriate messages. In [14] a real-time health status prediction system is proposed, this work focuses on applying machine learning especially decision tree on

978-1-5386-7850-3/19/$31.00 ©2019 IEEE

data streams received from socket streams using spark. Authors in [15] describe real-time flu and cancer surveillance system by mining twitter.

Most of this works involve machine learning, but in case of real-time machine learning applied to streaming data is not handled. On the other hand most of the healthcare analytics solution mainly focused on Hadoop which is a batch oriented computing. Heart disease is often the subject of researchers and doctors, both in diagnosis and treatment. Early detection of heart disease is fundamental for a rapid response and better chances of cure. Unfortunately, early detection of heart disease is often difficult because the symptoms of the disease at the beginning are absent. Yet, efficient and real-time system for predicting heart disease and patient consultation by doctors are not disposable because they require time, human resources, expensive material, knowledge and expertise. Based on available medical records data, early detection can be simplified by exploiting past cases to predict current situations based on machine learning and advanced generation processing engine for big data.

III. PROPOSED SYSTEM ARCHITECTURE

The purpose of this study is to develop a data processing, monitoring application. The system consists of two main sub parts, namely streaming processing and data storage and visualization. The first uses Spark MLlib with Spark streaming and applies machine learning model on health data events to predict heart disease. **Fig.1** shows the proposed model for real-time heart disease prediction and monitoring. Firstly, the data is sends from healthcare data sources.

A. Real time data processing

Spark streaming is a module supports scalable, fault-tolerant processing of live data streams. Incoming data stream is grouped into batches of interval (**Fig.2**) less than a second and processed by the batch processing spark engine, it can be processed using machine algorithms with high level function such as map and reduce. Finally, processed data may be pushed out to databases, file systems, and live dashboards for visualization and historical data analysis. Spark Streaming supports various input sources, including file based sources and network based sources such as receivers that communicate with socket based sources, the Twitter API stream, distributed stream and log transfer frameworks, such Flume, Kafka, and Amazon Kinesis.

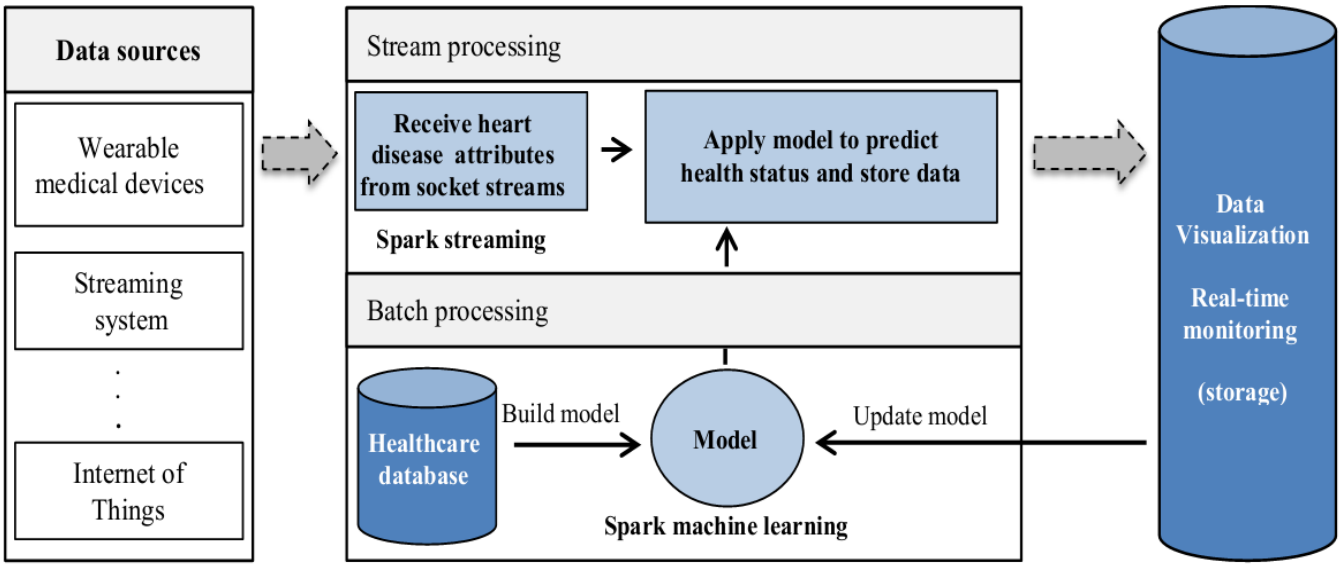

Fig. 2. Real-time heart disease prediction and monitoring overview

A DStream is defined by its input source and a time window called the batch interval.

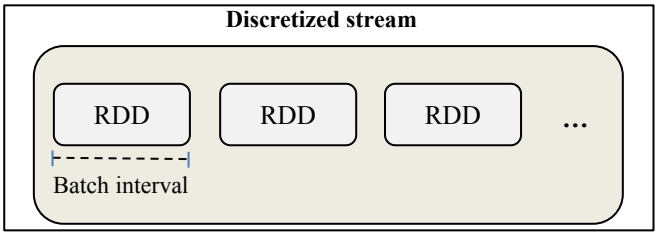

Fig.2. The discretized stream abstraction

B. Dataset

In this study, the dataset that is freely available and used in the majority of research papers is the heart disease dataset obtained from the UCI (University of California, Irvine C.A) has been used. The processed.cleveland.data of heart disease database [16] was used and analyzed. There are 303 records in this database. Each record in the database has fourteen attributes. The fourteen attributes are detailed in table below:

TABLE I. HEART DISEASE ATTRIBUTES

	Attributes	Description
1	Age	Age in years
2	Sex	Sex (1 = male, 0 = female)
3	Cp	Chest pain type (1= typical angina, 2= atypical angina, 3= non-anginal pain, 4= asymptomatic)
4	Trestbpss	Resting blood pressure (in mm Hg on admission to the hospital)
5	Chol	Serum cholestoral in mg/dl
6	Fbs	fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
7	Restecg	Resting electrocardiographic results
8	Thalach	Maximum heart rate achieved
9	Oldpeak	ST depression induced by exercise relative to rest
10	Exang	Exercise induced angina
11	Slope	The slope of the peak exercise ST segment
12	Ca	Number of major vessels (0-3)
13	Thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14	Num	Class (presence or absence of heart disease)

In this dataset, 139 (45.87%) records present presence of heart disease while 164 (54.13%) present absence of heart disease. The class label presents the presence of heart disease and absence of heart disease. In this part, the data is loaded form CSV file into an RDD.

As the focus of this paper is primarily on real-time processing, distributed and real-time classification and distributed storage, the database can be simply modified by other database. It just used to train the classification model.

C. *Data analytics*

Random forests are ensembles of decision trees and are one of the successful and efficient supervised classification algorithm that is capable of performing both classification and regression tasks. As the name implies, Random Forest builds the forest from several decision trees, in general the more trees in the forest the more robust the prediction and thus higher accuracy. Based on this, it has been selected to perform the prediction in the proposed system. To classify a new object based on attributes, the prediction of each tree is considered as a vote for one class. The label should be the class that receives the most votes. The good news is that random forest combines the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy. In this work, data events are performed using Spark streaming library and spark MLlib perform the Random forest implementation.

Based on the model error analysis, it has been discovered that the higher accuracy prediction stabilizes as the number of trees is between 4 and 10, and when the depth of decision tree is between 4 and 8.

- **Performance evaluation of random forest algorithm**

As stated in the previous discussion, machine learning model is tested on heart disease data set in the ratio 70-30. The diagnosis accuracy is maintained at 87.50%, which is calculated as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

- ***Confusion matrix***: confusion matrix describes performance of a classification model; it contains information about actual and predicted classifications performed by a classifier.

At the best accuracy of 87.50% which minimizes maxBins, maxDepth, numTrees parameters with Entropy impurity, the total test simple is equal to 88 and
True positive (TP) = 39 , True negative (TN) = 38
False Positive (FP) = 5 , False negative (FN) = 6
Sensitivity = 100 ∗ TP/ (TP+FN) = 100*39/45 = 86.66%
Specificity = 100 ∗ TN/ (FP+TN) = 100 ∗ 38/43 = 88.37%

TABLE II. CONFUSION MATRICE OF CLASSIFIER USED FOR CLASSIFICATION OF HEART DISEASE.

Predicted	Actual	
	Heart disease	No heart disease
Heart disease	39	5
No Heart disease	6	38

In order to show the efficiency of the proposed approach, other data records have been simulated and classified based on the model previously trained. **Fig.3** and **Fig.4** represents the performance evaluation of the implementation of random forest using spark MLlib and traditional tool of data mining.

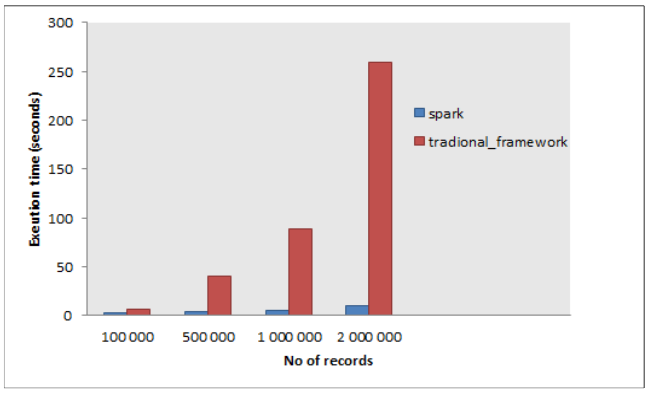

Fig.3. Time taken to build the random forest model

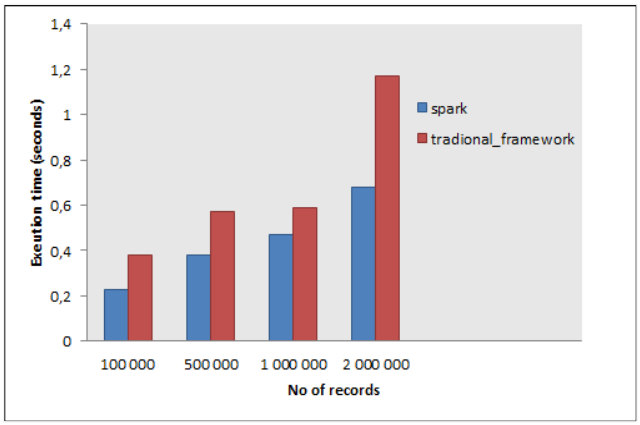

Fig.4. Time taken to test the random forest model

D. Data storage and visualization

The results as well as data streams generated by all the data sources needs to be stored in a distributed way to ensure the data availability with no single point of failure. Distributed databases are more scalable and provide better performance compared to traditional database systems.

Apache Cassandra [17] is a free, open-source, distributed NoSQL database system for managing large amounts of structured, semi-structured, and unstructured data. It is designed to manage very large amounts of data across many commodity servers; it provides high availability with no single point of failure. Here after data processing with spark, the result is stored in a table with a primary key through Cassandra. Data stored in database will be queried later for historical data analysis, visualizing, reporting and real-time monitoring.

IV. Results and Discussion

The proposed work is carried out in a single node cluster with core i7 processor having 8GB RAM in Linux platform through spark platform which integrates Random forest model with two stages, first involves analysis on healthcare dataset to build the machine learning model. The second uses the model in production to make predictions on live health data streams, heart disease observations is done throughout in single node cluster based on machine learning library MLlib, the computer-aided classification system was written using scala. Using the dataset above with varying maxDepth, maxBins and numTrees parameters different random forest models have been tested, the result is showing in Table 2. We simulated the multiple data streams by sending real-time data via simulated applications to the spark cluster. The implementation of use case is using Spark streaming written in Scala. Simulated applications generate more than 500000 data streams (heart disease attributes) per second.

V. Conclusion

This paper proposes a scalable system for heart disease monitoring using on Spark and Cassandra frameworks. This system focuses on applying real-time classification model on heart disease attributes for continuous monitoring of the patient's health. The system consists of two main sub parts, namely streaming processing and data storage and visualization. The first uses Spark MLlib with Spark streaming in which the classification model is performed by applying random forest to data events to predict heart disease. The seconds uses Apache Cassandra for storing the large volume of generated data. Once the data is streamed from all data sources, the proposed heart disease monitoring system is based on Spark framework and uses the random forest algorithm with MLlib for developing the prediction model to predict heart disease. Sensitivity, Specificity and Accuracy are calculated for evaluating of the prediction model. On the other hand system performance was performed using TCP messages of heart disease attributes. Integrating other big data technologies to our approach will be more efficient.

Developing a distributed and real-time healthcare analytics system using traditional analytical tools is extremely complex, while exploiting open source big data technologies can do it in a simpler and more effective way.

References

[1] A. Hazra, S. Mandal, A. Gupta, and A. Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review," Advances in Computational Sciences and Technology, 2017, 10, 2137-2159.

[2] Available from: http://hadoop.apache.org/ Online, accessed December 2017.

[3] D. Jeffrey, G. Sanjay, "MapReduce Simplified data processing on large clusters," Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation, (OSDI'04), Berkeley, CA, USA: USENIX Association, 2004, pp. 137–150.

[4] Available from: http://spark.apache.org/ Online, accessed December 2017.

[5] M. Zaharia, Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica,"Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," Technical Report UCB/EECS, EECS Department, University of California, Berkeley, 2011, pp. 2–2.

[6] Masethe, H.D., Masethe, M.A, "Prediction of heart disease using classification algorithms," In: World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, San Francisco, USA, 2014, 22–24 Oct.

[7] Rallapalli S., Gondkar R.R., Madhava Rao G.V, "Cloud Based K-Means Clustering Running as a MapReduce Job for Big Data Healthcare Analytics Using Apache Mahout," In: Advances in Intelligent Systems and Computing, 2016, vol 433. Springer, New Delhi.

[8] Goli-Malekabadi, Z., Sargolzaei-Javan, M., Akbari, M.K., "An effective model for store and retrieve big health data in cloud computing," Comput. Methods Programs Biomed, 2016, 132 ,75–82.

[9] Sarkar, Bidyut Biman, et al, "Personal Health Record Management System Using Hadoop Framework: An Application for Smarter Health Care," International Workshop Soft Computing Applications. Springer, Cham, 2016.

[10] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L, "Disease prediction by machine learning over big data from healthcare communities" IEEE Access, 2017, 5, 8869-8879.

[11] Rathore, M. M., Paul, A., Ahmad, A., Anisetti, M., and G. Jeon, "Hadoop-based intelligent care system (hics): Analytical approach for big data in iot," ACM Transactions on Internet Technology (TOIT),2017, 18(1):8.

[12] Akhtar, U., Asad, M., K.,, & Sungyoung, L. (2016). Challenges in Managing Real-Time Data in Health Information System (HIS). International Conference on Smart Homes and Health Telematics. Springer, Cham.

[13] Nair, Lekha R., Sujala D. Shetty, and Siddhanth D. Shetty, "Applying spark based machine learning model on streaming big data for health

status prediction," Computers & Electrical Engineering, 2018, 65 393-399.

[14] Ed-daoudy, A., Maalmi, K., "Application of machine learning model on streaming health data event in real-time to predict health status using spark" In: 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), IEEE, 2018 1-4.

[15] K. Lee, A. Ankit, C. Alok,"Real-time disease surveillance using twitter data: demonstration on flu and cancer," In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, 2013, pp. 1474-1477.

[16] Available from: https://archive.ics.uci.edu/ml/datasets/heart+Disease Online, accessed December 2017.

[17] Available from http://cassandra.apache.org Online, accessed December 2017