IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125110

Real-time machine learning for big data approach to early identification of heart disease

MOHANBABU.C[1], AMBIKA B[2]

Dept.of ECE, Chikkaballapur, SJC INSTITUTE TECHNOLOGY[1,2]

Abstract: The leading cause of death worldwide over the past few decades has been heart disease. Thus, regular monitoring and early detection of cardiac disease can lower the death rate. An vast amount of data has been continuously being generated by the exponential increase of data from various sources, including streaming systems, wearable sensor devices used in Internet of Things health monitoring, and others. A breakthrough in technology, streaming big data analytics and machine learning, has the potential to revolutionise the healthcare industry, particularly in the area of early heart disease detection. This technology might be more affordable and more potent. This research suggests a real-time cardiac disease prediction system built on Apache Spark to address this problem.

Keywords: big data, spark, distributed machine learning, heart disease, real-time

I. INTRODUCTION

The function of the heart, a muscle, is to pump blood throughout the body. It is the primary component of the body. One of the biggest health risks for males today is heart disease. The World Health Organisation (WHO) reports that heart attacks and strokes account for 80% of all fatalities worldwide. Therefore, having access to data and data mining techniques, particularly machine learning and early cardiac disease detection, can help patients prepare for a potential ailment. Data collection and processing are becoming increasingly complex in the healthcare industry as a result of the enormous volume of data (big data) generated from numerous sources, including streaming devices, advanced healthcare systems, high throughput instruments, sensor networks, the internet of things, and mobile applications.

The most popular framework for working with big data is Hadoop MapReduce, which is paired with Apache Spark, the next-generation big data processing engine. Hadoop MapReduce's main flaw is that it only enables batch processing; it is unsuitable for in-memory computing and real-time stream processing.

The MapReduce concept is expanded by Apache Spark to support more complex calculations. The core of Spark, which supports in-memory data storage and distributed computing, is a notion known as Resilient Distributed Datasets (RDDs). Spark Core and four libraries, including MLlib for machine learning and Spark Streaming for stream data processing, make up the current Spark project stack.

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

Big data analytics, particularly healthcare analytics, has recently been a significant topic for numerous studies. Machine learning has recently been used in healthcare by numerous researchers. To determine the optimum method of prediction, an experiment was conducted on the prediction of heart attacks. It has been suggested to employ cloud-based K-means clustering, which runs as a MapReduce task, to cluster healthcare data. For handling massive amounts of information, authors in have suggested a contemporary paradigm and system design. It is suggested to use Hadoop and HBase to create a web-enabled distributed and electronic health record personal health record management architecture. In order to forecast diseases using machine learning over large amounts of data from the healthcare industry, a multimodal disease risk prediction algorithm based on convolutional neural networks is used in.

A Hadoop-based intelligent care system is put forth in that exemplifies big data contextual sharing across all health system devices using the internet of things. In, a model for real-time analysis of large-scale medical data is put out. The method

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

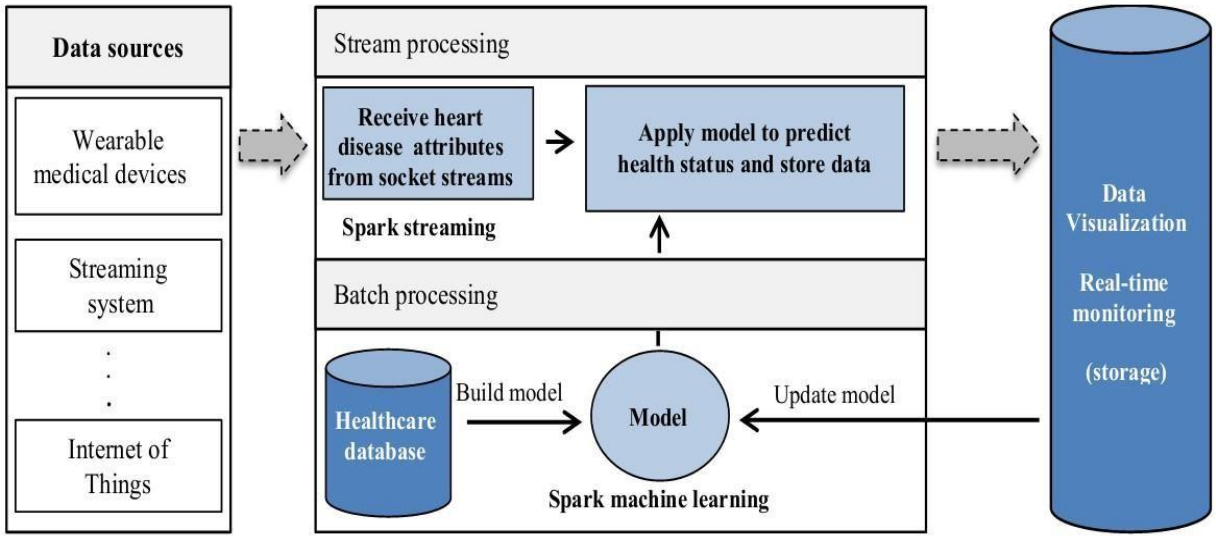
International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125110

is demonstrated by processing healthcare big data streams using Spark Streaming and Apache Kafka. On the other hand, several articles conduct stream computing over huge data. For instance, a prediction strategy is suggested in the suggested solution that is built on MLlib and a huge data processing engine. Twitter is used to collect and filter the data. The majority of these studies use machine learning, however they do not address real-time machine learning applied to streaming data. On the other hand, Hadoop, a batch-oriented computing platform, was the main focus of the majority of healthcare analytics solutions. Researchers and medical professionals frequently study heart disease, both for diagnosis and therapy. Early diagnosis of cardiac illness is essential for a quick reaction and higher recovery prospects. Unfortunately, because the illness's early signs are absent, it is frequently challenging to identify heart disease in its early stages. However, a reliable and timely method for forecasting cardiac disease and doctor-patient consultations are necessary since they demand time, people, expensive materials, knowledge, and resources. expertise. By using machine learning and an enhanced generation processing engine for big data, early detection may be made easier using the medical records data that is now available.

III. PROPOSED SYSTEM ARCHITECTURE



This study aims to develop a data processing system that includes streaming processing, data storage, and data visualisation. The first applies a machine learning model to health data events using Spark MLlib with Spark streaming to forecast cardiac disease. The proposed model for real-time cardiac disease monitoring and prediction is shown in Fig. 1. To begin with, the information comes from healthcare data sources.

A. Real-time data processing

Real-time data streams can be fault-tolerantly and scalably processed using a module called Spark streaming. The incoming data stream is split into intervals of less than a second by the batch processing technology used by the spark engine (Fig. 2). After that, high-level machine algorithms like map and reduce are used to process these batches. Finally, the data that has been processed could.

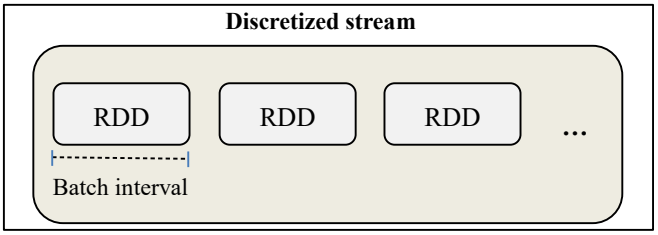


Figure 2. The abstraction of a discretized stream

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125110

B. Dataset

The heart disease dataset from the UCI (University of California, C.A.) has been used in this study. This dataset is freely available and used in the majority of research studies. We used and examined the processed.cleveland.data of heart disease database. This database contains 303 records. The database contains four attributes for each record. The table below provides more information on the 14 attributes: 139 individuals (45.87%) report having active cardiac disease, while 164 individuals (54.13%) report active disease.

C. Data analytics

Known as an effective and successful supervised classification technique that can handle both classification and regression tasks, random forests are collections of decision trees. As the name suggests, Random Forest creates the forest from a variety of decision trees; in general, the more trees in the forest, the more reliable and accurate the prediction will be. This has led to its selection to carry out the forecast in the suggested system. The forecast of each tree is taken into consideration as a vote for one class when categorising a new object based on attributes. The classification that receives the most votes ought to be the label. The good news is that random forest significantly improves on decision-making accuracy by combining flexibility with decision-tree simplicity.

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.

• Analysis of the random forest algorithm's performance

As was said in the preceding discussion, a machine learning model is tested using a data set on heart disease in a 70:30 ratio. The diagnosis accuracy is kept at 87.50% and is determined as follows:

• Confusion matrix: This matrix contains data on the actual and expected classifications made by a classifier, and it indicates the performance of a classification model.

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

IV. RESULTS AND DISCUSSION

The suggested work is executed on a single node cluster with a core i7 processor and 8GB RAM under Linux using the Spark platform, which incorporates the Random forest model with two stages, the first of which entails analysing a healthcare dataset to create the machine learning model. The second employs the model in production to produce predictions on real-time streams of health data; heart disease observations are made continuously in a single node cluster based on MLlib, and the computer-aided classification system was created in Scala. Different random forest models have been evaluated using the aforementioned dataset with varied maxDepth, maxBins, and numTrees parameters. The results are shown in Table 2. By feeding real-time data to the spark cluster via simulated applications, we were able to emulate the various data streams.

V. CONCLUSION

This study suggests a Spark and Cassandra-based scalable solution for heart disease monitoring. This method focuses on using a real-time classification model on characteristics of heart disease for ongoing patient health monitoring. The system is made up of two primary components: data storage and visualisation and streaming processing. In the first, heart illness is predicted using a classification model that applies random forest to data events using Spark MLlib and Spark streaming. The huge volume of created data is stored by the seconds using Apache Cassandra. The suggested heart disease monitoring system is based on the Spark framework and utilises the random forest method with MLlib to construct the prediction model to predict heart disease. Utilising open source big data technology can make developing a distributed and real-time healthcare analytics system simpler and more efficient than utilising conventional analytical tools.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125110

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 KMeans 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 realtime 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. 14741477.
[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