Proc. of the 2nd International Conference on Electrical, Communication and Computer Engineering (ICECCE)
12-13 June 2020, Istanbul, Turkey

Predictive Analytics Model Based on Multiclass Classification for Asthma Severity by Using Random Forest Algorithm

Wasif Akbar
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
sewasif@hotmail.com

Sehrish Saleem
Department of Computer Science
MNS University of Engineering and Technology Multan
Multan, Pakistan
sehrish@mnsuet.edu.pk

Wei-Ping Wu
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, China
wei-ping.wu@sipingsoft.com

Arslan Javed
Faculty of Electronic Information and Electrical Engineering School of Computer Science
Dalian University of Technology
Liaoning, China
arslan_ali_javed@yahoo.com

Muhammad Faheem
Department of Computer Science
National College of Business Administration & Economics
Lahore, Pakistan
faheem2725@gmail.com

Muhammad Asim Saleem
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
asim.saleem1@hotmail.com

***Abstract*—In modern life, health status prediction has become very crucial. Big data analysis plays a vital role to predict this perfectly. Asthma is a severe chronic disease with severe symptoms. Asthma disease is a chronic disease that leads to death. Researchers have focused on this for better decision making to predict the disease timely use of predictive analysis. This study proposes a Predictive Analytic Model for Asthma prediction using Random Forest (PAM-RF). Data of patients suffering from Asthma has been trained by a random forest approach which predicts to classify the data. Experiments are performed on Hadoop-spark which predicts the future health state of patients. The proposed approach has attained an accuracy of 98.80 percent to predict the asthma disease.**

Keywords—Data Analytics, Machine Leaning, Disease Prediction, Healthcare

I. INTRODUCTION

In the field of healthcare, the big amount of data is generated by multiple sources such as AHS (advance healthcare systems), streaming machines, IOT (internet of things), sensor networks, data processing, data collection and mobile applications [1]. Big data analytics plays a vital role in disease prediction. Predictive analytics allows researchers to establish predictive models that would not need numerous cases and also effective by the time [2][3]. Hadoop MapReduce [4][5] has been used to process big data analysis. But due to memory inconsistency, it only handles batch processing, therefore, it is not appropriate for real-time processing. To support more complex computations, Apache Spark expands the MapReduce pattern. Apache Spark [6] has been the most frequently used big data processing system. One of Hadoop MapReduce's main drawbacks is that it supports just batch processing. It is not appropriate for in-memory computation and real-time streaming. To support more complex computations, Apache Spark expands the MapReduce pattern. Spark provides a concept called (RDDs) Resilient Distributed Datasets [7] that is Spark's asthma designed to assist in-memory distributed computing and data storage. Spark software stack consists of four libraries, named as Spark streaming, MLlib, SQL and GraphX. This library's algorithms are designed to operate over a distributed dataset that is more appropriate for real-time analysis.

In this paper, we developed a Predictive Analytics Model for disease prediction using Random Forest (PAM-RF) to predict patients ' future status. We evaluate the efficiency of the proposed system for PAM-RF. There are five parts to the rest of the paper. Section 2 discussed related works literature. Section 3 describes the proposed PAM-RF scheme in which data is used to create the training model. Section 4 shows the effects of modeling and efficiency analysis of the proposed PAM-RF system in terms of execution and performance time and classification metrics. In the end, Section 5 reviews the findings and the course to be taken.

II. RELATED WORK

Sudha Ram et al. [8] presented a novel methodology using several data channels to estimate the number of patients to emergency department visits which are associated with asthma. This framework used data from Twitter and the environmental sensors to calculate the number of visits to the department for emergency asthma. Mayo Clinic's healthcare system was introduced by Dequan Chen et al. [9] to collect and store business data. This conducts clinical data for treatment, diagnosis, preventive measures, or clinical monitoring and non-clinical data for health informatics in medical research. In the Big Data platform, Alexandra et al. [10] discussed the complexities of machine learning techniques. Eventually, they concluded that in the age of Big Data Analytics, the machine learning strategies are perfectly suited to defined challenges. Min chenet et al. [11] introduced a risk analysis algorithm for healthcare data based on CNN multimodal disease. It includes an analysis of medical data for the detection of

978-1-7281-7116-6/20/$31.00 ©2020 IEEE

diseases in the healthcare community. Abdulsalam Yassine et al. [12] presented a model using smart home big data as a way of understanding and discovering patterns of human activity for applications in health care. They propose the use of regular pattern mining, cluster analysis, and prediction to quantify and interpret changes in energy usage induced by the behaviour of the occupants. Yichuan Wang et al. [13] developed a big data analytics system for healthcare sectors that defined five major data analytics technologies such as pattern analytics, unstructured data analytics, decisions support, predictive and traceability capabilities. Daniele Rav et al. [14] introduced deep learning applications in the field of bioinformatics, omnipresent computing and medical data processing and public health. Javier Andreu-Perez et al. [15] proposed new research theories for disease management from treatment to prevention for targeted treatment using big data technologies. Health data including computer imaging, health linguistics, sensor computing, and computational bioinformatics offer customized information from a variety of sources of data. H. Tamano et al. [16] presented a computing model for big-scale Hadoop ecosystems analysis. He used deep learning techniques for prediction. This takes full advantage of a cluster's parallel processing resources to process very large datasets efficiently in a manner that is fault-tolerant and scalable. A framework has been presented in [17] for real-time analysis of medical big data. The technique is illustrated by using Spark Streaming and Apache Kafka to process big data streams of healthcare. On the other side, several papers conduct stream computing over big data. For example, a predictive method is proposed in [18] the solution being proposed is based on the Big Data Processing Engine and MLlib. Twitter receives and filters the data, applies machine learning and sends suitable messages. A real-time health status analysis system has been proposed in [19] for data streams received from spark-based socket streams.

III. METHODOLOGY

The proposed PAM-RF (Predictive Analytics Model for disease prediction using Random Forest) method predicts a multiclass severity level of asthma disease. The asthma disease severity considered four classes such as Intermittent, Mild, Moderate and Severe. The proposed PAM-RF method used statistical classification technique to predict disease. Random Forest machine learning algorithm is used to build a training model with data. The dataset is collected from multiple hospitals in Pakistan. In Table I, we describe the attributes and their possible values in detail.

The proposed work uses the classification methodology of Random Forest to construct a prediction model of accuracy as this is one of the most widely used machine learning technique.

A. Dataset

The dataset contains 16000 patient's records which have different asthma severity and other respiratory diseases. The multiclass label shows the severity level of the disease. As the dataset in CSV file format, we are loaded into an RDD.

TABLE I. DESCRIPTION OF THE DATASET

Attribute	Description	Data Type	Domain of Values
age	Age of the patient	numerical	0-80
sob	Shortness of breath	numerical	2 Days to All Week Days
nta	Nighttime awakenings	numerical	0 Day to All Month Days
sufsc	SABA use for symptom control	numerical	2 Days to All Week Days
iwna	Interference with normal activity	nominal	None, Minor lim., Some lim., Extremely lim.
fev1pef	FEV1 (predicted)	numerical	60-80
fev1fvc	FEV1/FVC	numerical	60-85
severity	Severity	nominal	Intermittent, Mild, Moderate, Severe

B. Random Forest

The ensembles of decision trees are known as random forests. Random forests are one of the most efficient and successful supervised classification algorithms which are also able to perform regression and classification tasks. As the name indicates, from several decision trees Random Forest builds the forest, usually the more trees in the forest the more reliable prediction and precision. Because of this, the prediction in the proposed system has selected the prediction of each tree is known as a vote for one class to identify a new object depending on attributes. The label must be the most votes receiving class. The random forest combines decision trees ' simplicity with adaptability resulting in a massive improvement of accuracy. performance of the model.

The confusion matrix characterizes a classification model's efficiency; it includes information about a classifier's actual and predicted classifications.

TABLE II. CONFUSION MATRIX OF RADOM FOREST ALGORITHM

Predicted	Actual			
	Intermittent	**Mild**	**Moderate**	**Severe**
Intermittent	1510	0	1	83
Mild	0	1651	0	0
Moderate	0	0	1537	0
Severe	78	0	0	1554

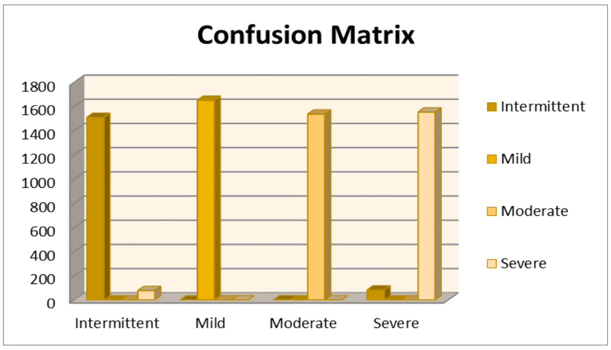

Fig. 1. Asthma severity comparison of the confusion matrix

The Table II shows the number of classified instances of given dataset according to its class. Intermittent class has 1510 classified instance.1 Moderate 83 Sever class instance that are actually relates in Intermittent class. Mild class has 1651 and moderate has 1537 instances. Sever class has 1554 instances. 78 instances are intermittent class but these instances are belonging to sever class. Figure 1 displays the graphical representation of the confusion matrix, Maximum instances are related to Mild and minimum are related to the intermittent class.

TABLE III. EVALUATION METRICS-RDD-BASED BINARY CLASSIFICATION SPARK MLIB

Evaluation Metrics-RDD-Based Binary Classification SPARK Mlib			
Parameters	**Precision**	**Recall**	**F1-Score**
Class-0 (Intermittent)	0.9508	0.9473	0.9490
Class-1 (Mild)	1	1	1
Class-2 (Moderate)	0.9993	1	0.9996
Class-3 (Severe)	0.9492	0.9522	0.9507
Weighted	0.9747	0.9747	0.9747

Table III deals with the evaluation metrics of Apache Spark Mlib (Machine Learning Library). Used machine learning algorithms for data prediction and the evaluation of prediction model Apache Spark provide RDD-based API. Binary Classification separates data into multiclass classification. There are 4 classes for the predication of data and evaluation by the Binary Classification using multiclass classification. Each class (Intermittent, Mild, Moderate, and Severe) has its parameters like Precision, Recall, and F1 score and the weighted values of these parameters are show the overall performance of the prediction model.

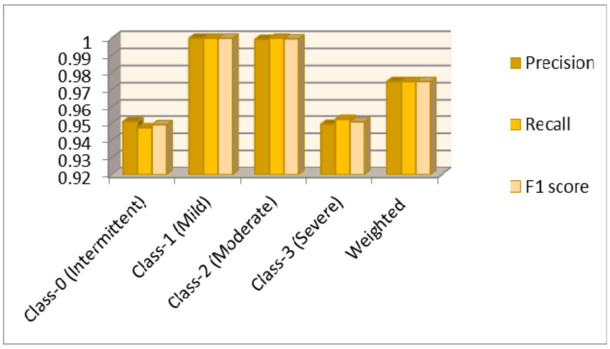

Fig. 2. Evaluation matrix of multiclass asthma severity

Figure 2 represents the prediction classes that have evaluation matrixes. Class-0(Intermittent) has minimum precision, recall, and F1 score and Class-1(Mild) has maximum evolution parameters.

The proposed system is tested on Ci3 1.7 Ghz 64-bit processor with 4GB RAM with window 10 OS with Netbeans, Java and Apache Spark and Intel Xeon-E5520, Cache 256 MB (With 8 Cores), 8 GB RAM, 64-bit Window Server 2012 through the spark framework that implements a two-stage random forest model, analysing the healthcare dataset first involves building a machine learning model. The second applies the method in development to predict live health data sources, measurements of heart disease are made in a single node cluster based on the MLlib library of machine learning, Java has been used to write the computer-aided classification system. Table 3 defines the methods for evaluating the proposed system.

TABLE IV. PREDICTION MODEL EXECUTION ON APACHE SPARK IN MILI SECONDS

Prediction Model Execution on Apache Spark in Mili Sec.		
Parameters	**Core i3**	**Xeon-E5520**
Spark Initialization Time	35	14
Training Model by Data	17	17
Evaluation of Data	22	14
Prediction of Test Data	20	7
Accuracy of Data Classification	97.4	97.4
Accuracy of Prediction	98.8	98.8

Table IV contains the execution time of the prediction model with data evaluation and prediction of testing data. Firstly, the Apache Spark initialization than provides the training data to the model by splitting the given dataset into training and testing. After the train, the prediction model Evaluation Matrix is performed on the given dataset for the classification using Apache Spark Evaluation Matrix RDD Based API by Multiclass Binary Classification. After that, the test data are given to model for the prediction than the model calculated the data classification and prediction accuracy.

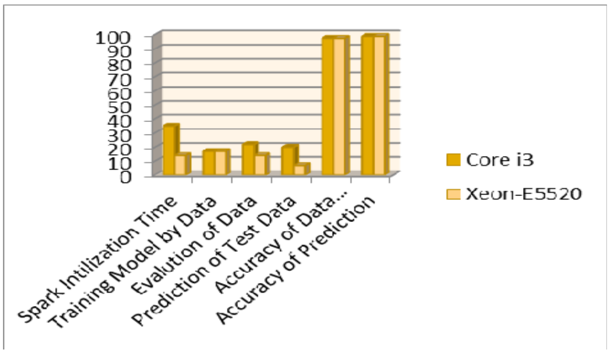

Fig. 3. System execution time graph

Figure 3 display the graphical representation of the execution time on two systems. Maximum time has taken by the Apache Spark initialization in both systems. Total time has taken by the prediction model 1.5 Seconds on the Core i3 system and 55 Seconds on the Server system. Classification accuracy is 97.4% using Multiclass Binary Classification and Prediction accuracy is 98.80% by using the Random Forest Algorithm of Machine Learning.

IV. CONCLUSION

In this paper, we propose a random forest method to construct a classifier model that predicts the future health of asthma disease. We used a map-reduce technique for random forest technology that integrates with the Apache Spark framework to perform predictive analysis of big data, which helps reduce computational complexity due to its parallelism. The proposed PAM-RF scheme can effectively classify and predict future status from asthma datasets. Finally, the hospital's dataset is simulated, and future predictions ensure that the proposed system has improved significantly in terms of processing time, CPU utilization, and accuracy.

REFERENCES

[1] L. Wang and C. A. Alexander, "Big Data in Medical Applications and Health Care," 2015.

[2] I. Palit and C. K. Reddy, "Scalable and Parallel Boosting with MapReduce," vol. 24, no. 10, pp. 1904–1916, 2012.

[3] P. K. Sahoo, S. K. Mohapatra, and S. Wu, "Analyzing Healthcare Big Data with Prediction for Future Health Condition," vol. 3536, no. c, pp. 1–13, 2016.

[4] Available from: http://hadoop.apache.org/ Online, accessed December 2019.

[5] J. Dean and S. Ghemawat, "MapReduce : Simplified Data Processing on Large Clusters," pp. 1–13.

[6] Available from: http://spark.apache.org/ Online, accessed December 2019.

[7] M. Zaharia et al., "Resilient Distributed Datasets : A Fault-Tolerant Abstraction for In-Memory Cluster Computing."

[8] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting Asthma-Related Emergency Department Visits Using Big Data," vol. 2194, no. c, 2015.

[9] D. Chen et al., "Real-Time or Near Real-Time Persisting Daily Healthcare Data Into HDFS and ElasticSearch Index Inside a Big Data Platform," IEEE Trans. Ind. Informatics, vol. 13, no. 2, pp. 595–606, 2017.

[10] A. L. Heureux and G. S. Member, "Machine Learning With Big Data : Challenges and Approaches," IEEE Access, vol. 5, pp. 7776–7797, 2017.

[11] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," vol. 3536, no. c, pp. 1–10, 2017.

[12] A. Yassine, S. Singh, and A. A. Member, "Mining Human Activity Patterns from Smart Home Big Data for Healthcare Applications," vol. 3536, no. c, pp. 1–10, 2017.

[13] Y. Wang, L. Kung, and T. Anthony, "Technological Forecasting & Social Change Big data analytics : Understanding its capabilities and potential bene fi ts for healthcare organizations," Technol. Forecast. Soc. Chang., vol. 126, pp. 3–13, 2018.

[14] D. Rav, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-perez, and B. Lo, "Deep Learning for Health Informatics," vol. 21, no. 1, pp. 4–21, 2017.

[15] N. Mehta and A. Pandit, "International Journal of Medical Informatics Concurrence of big data analytics and healthcare : A systematic review," Int. J. Med. Inform., vol. 114, no. March, pp. 57–65, 2018.

[16] M. Viceconti, P. Hunter, and R. Hose, "B ig data , big knowledge : big data for personalised healthcare," vol. 2194, no. c, 2015.

[17] U. Akhtar, A. M. Khattak, and S. L. B, "Challenges in Managing Real-Time Data in Health Information System (HIS)," vol. 1, pp. 305–313, 2016.

[18] L. R. Nair, S. D. Shetty, and S. D. Shetty, "big data for health status prediction R," vol. 0, pp. 1–7, 2017.

[19] A. Ed-daoudy, K. Maalmi, U. Sidi, and M. Ben, "Application of machine learning model on streaming health data event in real-time to predict health status using Spark," 2018 Int. Symp. Adv. Electr. Commun. Technol., pp. 1–4, 2018.