IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2016, accepted December 22, 2016, date of publication January 4, 2017, date of current version January 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2647619

Analyzing Healthcare Big Data With Prediction for Future Health Condition

PRASAN KUMAR SAHOO[1,2], (Senior Member, IEEE), SUVENDU KUMAR MOHAPATRA[4], AND SHIH-LIN WU[1,2,3], (Member, IEEE)

[1]Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, 33302, Taiwan
[2]Department of Cardiology, Chang Gung Memorial Hospital, Taoyuan, 33305, Taiwan
[3]Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City, 24301, Taiwan
[4]Department of Electrical Engineering, Division of Computer Science and Information Engineering, Chang Gung University, Taoyuan, 33302, Taiwan

Corresponding author: S.-L. Wu (slwu@mail.cgu.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant 105-2221-E-182-050 and Grant 105-2221-E-182-043.

ABSTRACT In healthcare management, a large volume of multi-structured patient data is generated from the clinical reports, doctor's notes, and wearable body sensors. The analysis of healthcare parameters and the prediction of the subsequent future health conditions are still in the informative stage. A cloud-enabled big data analytic platform is the best way to analyze the structured and unstructured data generated from healthcare management systems. In this paper, a probabilistic data collection mechanism is designed and the correlation analysis of those collected data is performed. Finally, a stochastic prediction model is designed to foresee the future health condition of the most correlated patients based on their current health status. Performance evaluation of the proposed protocols is realized through extensive simulations in the cloud environment, which gives about 98% accuracy of prediction, and maintains 90% of CPU and bandwidth utilization to reduce the analysis time.

INDEX TERMS Big data, cloud, healthcare, prediction.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) and mobile networks allow in-hospital and outdoor patients monitoring through Internet of Things (IoT) [1] in which patients are equipped with different smart devices such as in-plant pacemaker, Electrocardiogram (ECG), Electromyography (EMG), Electroencephalography (EEG) and motion sensors, etc. These wearable devices collect health related data such as body temperature, blood pressure and heart rate, which can be applied in physical fitness tracking and medical treatment. Big data in healthcare [2] is an analytic environment to handle the massive volume of structured and unstructured patient data. According to the analysts, the healthcare data volume of USA healthcare system has reached to 150 exabytes in 2011 [3] and has increased to zettabyte scale [4] in the current time. Similarly, the California-based health network Kaiser Permanente has 9 million members and the data [3] collected from Electronic Health Records (EHRs) including doctor notes, clinical reports and pathological images range from 26.5 to 44 petabytes.

The health data are attributed as big data, which is defined by *5Vs* in terms of Volume, Velocity, Variety, Value, and Veracity. The collected patient data are of peta or zeta bytes, which describe the volume. The velocity is expressed in terms of data arrival rate from the patients. Variety explains the diversified data sets with respect to the structured, semi-structured and unstructured data sets such as clinical reports, EHRs, and radiological images and veracity explains the truthfulness of the data sets with respect to data availability and authenticity. The collected data are transformed into meaningful insights, which explain the value in 5Vs.

Physiological data of patients are the primary and vital entities in healthcare big data analytic. Hence, valid raw data must be collected with an efficient manner in a medical environment. In advanced healthcare systems, the patient data are collected [5] through wearable devices equipped with different types of sensors. Recently, the advancement in mobile devices [6] such as multi-sensor equipped smart phones are also used as the data collection devices. Hence, colossal amounts of patient data are generated within a hospital network, which needs to be stored and analyzed efficiently. Therefore, a cloud computing [7] enabled distributed storage and processing environment is essential to store and process

the healthcare data, which can be accessed anywhere and anytime.

Now-a-days various data intensive applications are emerged, which need some efficient analytic models. Many stochastic approaches [8] are considered by different authors in the recent past for healthcare parameter analysis. Moreover, the similarity [9] between health parameters of a patient is considered by the physicians for better decisions. Big data analytic is applied in healthcare [10] to identify the clusters of patients, diseases and future predictions with the help of various machine learning tools [11]. In a learning healthcare system [12], data are analyzed and used as insights continuously for patient care. During this process, the patient data are combined with the clinical reports for better suggestions and decisions.

So far limited analyses have accomplished among the patients taking different numbers of health parameters of same or different departments. Even, the existing models cannot support both analysis and processing for the large volume of multi-structured healthcare data. Recently, the high performance of cloud platform provides a scalable and distributable parallel processing framework, i.e. MapReduce [13] for healthcare data processing. MapReduce has the capability to process the large volume of data in parallel on a cloud. The major benefits of MapReduce framework are the scalability and fault-tolerance during massive data processing on a large cloud. Hence, a hybrid model of stochastic and parallel processing framework is planned in a medical environment to process and analyze the huge volume of healthcare data. In our work, MapReduce parallel processing framework [14] is used as a backbone for healthcare big data analysis. Further, the proposed work is extended to a prognosis model for future health condition prediction of the patients.

Rest of the paper is organized as follows. Existing works on big data in the cloud environment are discussed in Section II. Section III describes the system model of our work. A probabilistic data collection model is designed in Section IV. The healthcare data analysis and processing mechanisms are given in Section V. The prediction of a patient future health condition is designed in Section VI. Performance evaluation of our proposed models is given in Section VII and concluding remarks are made in Section VIII.

II. RELATED WORKS

How to analyze data to derive meaning information is highly essential for studying the mammoth health related raw data and to predict the future health condition. The temporal and spatial correlations of sensing pilate exercise data are analyzed in [15] for knowing the relief of lower back pain by keeping track of the patient's body motions. However, limited works are performed on correlation analysis of healthcare parameters among different patients. In [16], in-hospital mortality of emergency department patients is predicted using a local big data-driven random forest model. However, only clinical data of patients are considered in the existing models ignoring the history of disease symptoms. A brief survey is performed on advantages and disadvantages of applications and technical requirements for in-hospital and BAN patients monitoring in [17]. Hu *et al.* [18] have designed the data acquisition mechanism by using sensors, log files and web crawler in various applications. However, the frequency of the patient visit is not considered during data collection.

A new big data architecture with methodology for healthcare is proposed in [3] and Zhang *et al.* [19] propose a task-level adaptive and scalable MapReduce framework, which can estimate the future arrival rate of workload on the map and reduce phases. In another prospective, MapReduce framework is designed to reduce the re-computation for incremental iterative computations in [20]. An online community-based health services is proposed in [21], where the health data are collected and mined through some questionnaires and their respective answers. A scalable and distributable method is proposed in [22] to find the similarity among patients by modifying the MapReduce framework. This method can support the storage and information retrieval over the time stamp. However, the visiting frequency, health parameters and hidden symptoms of patients are very important but are not taken into consideration for analyzing and processing the data in this work.

Future disease prediction is very crucial and important for the patients with chronic diseases. Many disease prediction models have been proposed in the recent past. In [23], different types of artificial neural network (ANN) techniques are discussed for disease prediction. However, ANN takes longer time for training the model due to diversified weights associated with each layer. Even, any small change in the input data set affects the model, which gives unstable output. In [24], the feature stability is observed by using the regression based feature graph for the clinical prognosis in high-dimensional electronic medical records. A predictive framework is designed in [25] to integrate the EHR data with risk factors to effectively predict the osteoporosis and bone fractures. Henriques *et al.* [26] predict the decompensation of heart failure by considering the physiological data of the patients. However, the hidden symptoms of the diseases are not considered in the current prediction models.

As given in [27], bio-sensors such as ECG, EMG and EEG are used to collect and transmit the health parameters to back-end servers for processing. However, the visiting frequency of the patient with respect to the doctor and department in a hospital is not considered, which has major impact on the data collection process. Though many researchers have proposed the deployment and sensing strategies of the body sensors for collecting data, none of them have developed the data collection models of the indoor and outdoor patients based on the frequency of visits to a hospital. In our proposed work, data collection models are developed taking physiological parameters and hidden symptoms of the diseases of the patients. Furthermore, the correlation analysis is incorporated with disease prediction among the patients in a hospital. Hence, the main objectives of this paper are to find the distinguished

characteristics of the diseases by introducing the correlation analysis of healthcare parameters, which can be summarized as follows.

- Propose a probabilistic data collection model based on the frequency of out-patient's visits and volume of data generated from the patients with BAN.
- Correlation analysis algorithms are designed for the patients of intra and inter departments of the hospitals.
- An algorithm for predicting future health condition of patients based on their current health status is designed.

III. PROBLEM FORMULATION

Consider a cloud based healthcare environment with h numbers of hospitals in a set $H = \{H_1, H_2, \ldots, H_h\}$, where $h \in H$ as shown in Fig. 1. Let various departments be associated with one hospital and for simplicity, it is assumed that same and equal numbers of departments are present in each hospital. Let, $DP = \{DP_1, DP_2, \ldots, DP_\delta\}$ be the set of δ numbers of departments associated with each hospital. Besides, each department is coupled with different numbers of doctors, out-patients and BAN patients, which are the sources for generating the big data. It is to be noted that out-patients are the patients who visit the hospital for treatment without staying there overnight. BAN patients are the chronic patients fitted with smart body sensors to monitor their health condition round the clock. For simplicity, throughout the paper, we refer to the out-patients and BAN patients as patients and BAN, respectively.

Let, d be the numbers of doctors present in a set D^k_{ij}, where $j = \{1, 2, \ldots, d\}$ in the i^{th} department of k^{th} hospital, $\forall i \in DP$ and $\forall k \in H$. Thus, $D^k_{ij} = \{D^k_{1d} \cup D^k_{2d} \cup \ldots \cup D^k_{\delta d}\}$, $\forall i \in DP$, $\forall k \in H$. For example, D^3_{21} represents the doctor 1 that belongs to the department 2 in hospital 3. Let, P^k_{ij} be the set of patients, where $j = \{1, 2, \ldots, p\}$ in i^{th} department of k^{th} hospital, $\forall i \in DP$ and $\forall k \in H$. Hence, p numbers of patients are present in the i^{th} department of k^{th} hospital. Therefore, $P^k_{ij} = \{P^k_{1p} \cup P^k_{2p} \cup \ldots \cup P^k_{\delta p}\}$, $\forall i \in DP, \forall k \in H$. For example, P^3_{21} represents the patient 1, which belongs to the department 2 in hospital 3. It is assumed that patients with BANs are also admitted to a hospital, which could be either a patient or a BAN at a time. Similarly, let b be the number of BANs present in a set B^k_{ij}, where $B^k_{ij} = \{B^k_{1b} \cup B^k_{2b} \cup \ldots \cup B^k_{\delta b}\}$, $\forall i \in DP, \forall k \in H$ and different number of BANs are available in various departments within a hospital. For example, B^3_{21} represents the BAN 1 that belongs to the department 2 in hospital 3.

In our proposed model, a window based temporal data collection and monitoring model is used to enhance the quality of patient monitoring. Let, $T = \{0, 1, 2, \ldots, t\}$ be a continuous time frame, which is divided into w number of windows, where each window consists of z units of time duration. Each time duration could be considered as a minute, an hour, a week, a month or a year that depends on the applications. Accordingly, $D^k_{ij}(w)$, $P^k_{ij}(w)$, and $B^k_{ij}(w)$ represent the volume of data generated from the doctors, patients and BAN, respectively in each window w. The collected data within window w are stored in different cloud data centers as shown in Fig. 1. Let, $\{DC_1, DC_2, \ldots, DC_n\}$ be the N numbers of geo-distributed data centers located in the cloud, where $n \in N$. These data centers are connected through M numbers of gateways $G = \{GW_1, GW_2, \ldots, GW_m\}$, where $m \in M$. In our framework, H numbers of those hospitals are connected with those N numbers of geo-distributed data centers via M numbers of gateways.

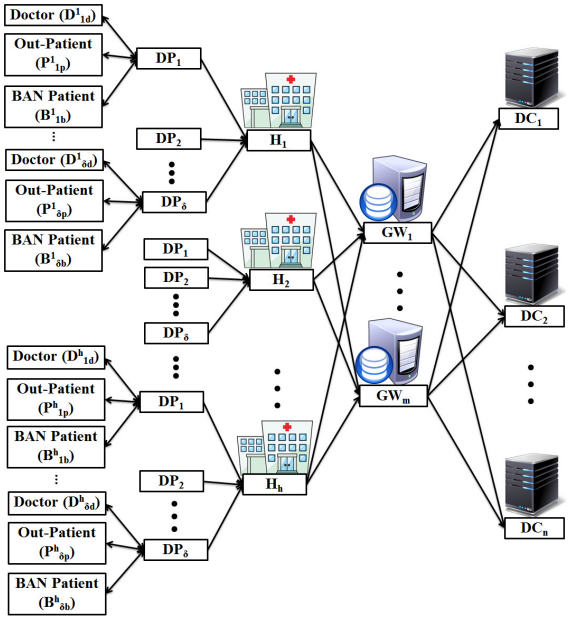

FIGURE 1. Communication model between hospital and cloud data centers.

IV. DATA COLLECTION MODEL

In traditional healthcare systems, the patient data are collected, stored and analyzed in a traditional manner, which cannot support the diagnosis of complex health conditions. However, in our proposed data collection scheme, doctors, patients and BANs are considered altogether as the sources of generating data based on the frequency of visits (f) of a patient instead of considering only numbers of patients as analyzed in traditional schemes. A window based [28] temporary data collection and monitoring models are used to enhance the usefulness of patient monitoring. The window size could be modified as per the health problem or requirements. Mostly, this is more beneficial for the self-monitoring and time series patients, where the patient condition is observed based on the series of health related parameters. For example, blood pressure data of patient A is recorded on the time frame $T = \{0, 1, \ldots, t\}$ with window frames $W = \{W_1, W_2, \ldots, W_w\}$, as shown in Fig. 2, where the duration of each window W_i contains 3 *units*. In the second window (W_2), the average blood pressure data ($Avg.W_2$) of three time slots (4, 5, 6) is recorded as 140, which is greater than the critical condition ($if\,(Avg.W_w > 130)$). Those recorded data with time series are collected by using

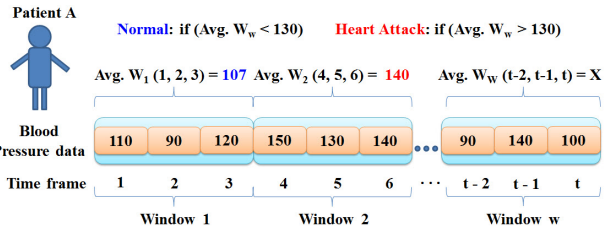

FIGURE 2. Window based data collection and monitoring.

our proposed data acquisition scheme and is transmitted to the data centers in the cloud for storage and analysis.

In healthcare big data environment, the physiological data, EHR, 3D imaging, radiology images, genomic sequencing, clinical, and billing data are the sources of big data, which describe the volume. Real-time and emergency patient monitoring such as BAN patients, heart beat monitoring and Intensive Care Unit (ICU) patient monitoring are the sources of streaming data, which describe the velocity of the data. Similarly, the healthcare data such as ECG, EMG and clinical reports are the the unstructured data, whereas the patients visits, personal records are the structured data, which describe the verity. Though most of the papers consider the patient's physiological data as the big data [3], we include visiting frequency of the patients to the hospitals in our big data processing models.

A. DATA ACQUISITION SCHEME

The numbers of visits of a patient to consult the doctors in different hospitals need to be analyzed as they can generate data in each visit. Without loss of generality, let us assume that a patient (P_{ij}^k) visits f times to a department (DP_i) within w intervals. During the visit of each patient, let $\wp_f$, $\wp_{DP}$ and $\wp_D$ be the probabilities of patients visiting frequency of hospitals, departments, and doctors in the window w, respectively. It is to be noted that $\wp_V$ and $\wp_{BA}$ are the least value of the probability of a patient that visits the hospital and BANs, respectively.

Theorem 1: Probability of visit $\wp_V$ of a patient to a hospital is at least $\frac{f}{d\delta w}$.

Proof: Let, f be the frequency of visits by a patient P_{ij}^k to the i^{th} department of k^{th} hospital, where $i \in DP$ and $k \in H$. Hence, the probability of frequency of visits $\wp_f$ of a patient to the hospital within the window w can be expressed as $\frac{f}{w}$. Similarly, the visiting probability to a department $\wp_{DP}$ of a patient within the hospital is $1/\sum_{\imath=1}^{\delta} DP_i$. Probability of a patient consulted by a doctor within one department $\wp_D = 1/\sum_{\imath=1}^{d} D_i$. If there are d numbers of doctors present in δ numbers of departments in a hospital, the total probability of visits of a patient can be expressed as $\wp_V = \wp_f * \wp_{DP} * \wp_D$. Further, $\wp_V$ can be $\frac{f}{w} * \frac{1}{\sum_{\imath=1}^{\delta} DP_i} * \frac{1}{\sum_{\imath=1}^{d} D_i}$. If we proceed further, $\wp_V(w)$ becomes $\frac{f}{d\delta w}$. ■

It is to be noted that probability of visits of a patient increases monotonically with f and w. In another scenario, a BAN (B_{ij}^k) is associated with the department in a hospital, which generates data with time. Similarly, the indoor patients also generate data time to time. In both scenarios, probability of frequency of visits ($\wp_f$) is set to be 1, as the BAN or indoor patients can generate data throughout the observed time (w).

Theorem 2: Probability of consultation of a BAN $\wp_{BA}$ or indoor patients in a department is at least $\frac{1}{d\delta}$.

Proof: Let, B_{ij}^k be the j^{th} BAN associated with i^{th} department of k^{th} hospital, $\forall i \in DP$ and $\forall k \in H$. If d and δ are the number of doctors and departments present in a hospital, respectively, the visiting probability of a patient to the doctors and to a department are $\wp_D$ and $\wp_{DP}$, respectively. The total probability can be expressed as $\wp_{BA} = 1 * \wp_{DP} * \wp_D$. Hence, $\wp_{BA} = 1 * \frac{1}{\sum_{\imath=1}^{\delta} DP_i} * \frac{1}{\sum_{\imath=1}^{d} D_i}$. Finally, $\wp_{BA}$ can be obtained as $\frac{1}{d\delta}$. ■

It is observed that the probability $\wp_{BA}$ increases if the BANs associated with multiple doctors and departments increase. As given in [29], the clinical test data and radiological images are considered as structured and unstructured data sets, respectively. According to the authors, one terabyte of clinical text data and 19 terabytes of image data are generated by 250000 patients per year. In our data acquisition scheme, both text and image data of the patients are also considered, which are generated during visits of the patients. Let u and v megabytes be the size of each text and image data, respectively and φ^p be the amount of data generated by a patient p during a single visit. Thus, φ^p is the aggregated amount of both text (φ_{TD}^p) and image (φ_{ID}^p) data of a patient p, which can be expressed in Eq. (1).

$$\varphi^p(w) = \sum_{\imath=1}^{x} \varphi_{TD_\imath}^p(w) * u + \sum_{\jmath=1}^{y} \varphi_{ID_\jmath}^p(w) * v \tag{1}$$

Where, φ^p(w) is the data generated in the window w. In Eq. (1) and (2), x and y are the numbers of generated text and image documents, respectively for a patient p in a single consultation window w. Similarly, φ^b is the aggregated amount of text (φ_{TD}^b) and image (φ_{ID}^b) data generated from a BAN b.

$$\varphi^b(w) = \sum_{\imath=1}^{x} \varphi_{TD_\imath}^b(w) * u + \sum_{\jmath=1}^{y} \varphi_{ID_\jmath}^b(w) * v \tag{2}$$

Considering the visiting probability of a patient within a window w, amount of data generated from a patient $\Phi^p(w)$ can be expressed as given in Eq. (3).

$$\Phi^p(w) = \varphi^p(w) * \wp_V(w) \tag{3}$$

The total amount of data $\Phi^b(w)$ generated by a BAN b within a window w is given in Eq. (4).

$$\Phi^b(w) = \varphi^b(w) * \wp_{BA}(w) \tag{4}$$

Subsequently, Φ^δ(w) is the amount of data collected from a department δ within a window w, which can be expressed

in Eq. (5).

$$\Phi^{\delta}(w) = \Phi^{p}(w) + \Phi^{b}(w) \tag{5}$$

The accumulated data $\Phi^{i}_{Tot}(w)$ of the patients and BANs collected within the window w in the i^{th} hospital can be represented as given in Eq. (6), where $i \in H$.

$$\Phi^{i}_{Tot}(w) = \sum_{\iota=1}^{\delta} \Phi^{\iota}(w) \tag{6}$$

Since, $h \in H$ is the number of hospitals present in the healthcare system, the total amount of data $\Phi^{h}_{Tot}(w)$ can be expressed as given in Eq. (7).

$$\Phi^{h}_{Tot}(w) = \sum_{\jmath=1}^{h} \sum_{\iota=1}^{\delta} \Phi^{\jmath\iota}(w) \tag{7}$$

B. TASK ALLOCATION MODEL

In our task allocation model, the tasks are assigned with respect to the computation (μ) rate of an active server and types (τ) of the tasks. It is assumed that the tasks are processed instantly without any buffering or queuing delay in the data centers. Before execution of the tasks, the data are distributed [30] with multiple replicas among the active servers (θ) of the data centers for load balancing and data locality purpose. In our proposed system, arrival of tasks is considered in two phases, i.e. from the hospital to the gateway and gateway to the data center's active servers. Let, $\lambda_{h,g}(w)$ be the arrival rate of the tasks from a hospital $h \in H$ to the gateway $g \in G$ and $\lambda_{g,\theta}(w)$ be the arrival rate of tasks from a gateway $g \in G$ to the active servers θ of a data center in a window w. All incoming tasks from different hospitals to the active servers of the cloud data centers via gateways are processed and expressed as given in Eq. 8.

$$\sum_{\iota=1}^{h} \lambda_{\iota,g}(w) = \sum_{\jmath=1}^{g} \lambda_{\jmath,\theta}(w) = \sum_{\ell=1}^{\theta} \mu_{\ell}(w) \tag{8}$$

Further, the tasks are categorized into two types based on the processing time and priority. In healthcare system, some tasks may need short processing time such as doctor's query and short-term analysis. Some other tasks may need long processing time such as data backup, migration, and integration. Similarly, priority tasks are defined as the queries coming in emergency situations such as queries during any surgery. Let, $\tau^{h}_{s}(w)$ and $\tau^{h}_{l}(w)$ be the short and long processing type of the tasks during a window w in a hospital $h \in H$, respectively. Similarly, let $\tau^{h}_{pr}(w)$ and $\tau^{h}_{po}(w)$ be the amount of priority and posteriority tasks available during a window w for processing in a hospital $h \in H$, where $\tau^{h}_{s} = \tau^{h}_{pr} + \tau^{h}_{po}$. The arrival rate $\lambda(w)$ is bounded by the task $\tau(w)$ during each window w as given in Eq. 9.

$$\left.\begin{aligned} \tau^{h}_{s}(w) + \tau^{h}_{l}(w) &= \sum_{\iota=1}^{h} \lambda_{\iota,g}(w) = \sum_{\jmath=1}^{g} \lambda_{\jmath,\theta}(w) = \sum_{\ell=1}^{\theta} \mu_{\ell}(w) \\ \tau^{h}_{pr} + \tau^{h}_{po} + \tau^{h}_{l}(w) &= \sum_{\iota=1}^{h} \lambda_{\iota,g}(w) = \sum_{\jmath=1}^{g} \lambda_{\jmath,\theta}(w) = \sum_{\ell=1}^{\theta} \mu_{\ell}(w) \end{aligned}\right\} \tag{9}$$

V. DATA ANALYSIS MODEL

As discussed in the data acquisition phase, Φ^{h}_{Tot} amount of data are collected for analysis. Those bulky data sets are divided into small number of chunks for parallel processing. Let, χ be the size of each partition (or chunk) of the input data sets. Now, the total number of equal partitions can be $U_p = \lfloor \frac{\Phi^{h}_{Tot}}{\chi} \rfloor$. MapReduce model is used to process those huge amounts of healthcare data, which is discussed in the following subsection.

A. ANALYSIS IN MapReduce FRAMEWORK

In the data input phase of the MapReduce framework [14], [30], [31], U_p quantity of data blocks are induced, which are collected from the patients and the BANs in the data acquisition phase. Those U_p number of data blocks are distributed among the active servers to achieve the data locality before any task execution. Let, $Q = \{Q_1, Q_2, \ldots, Q_q\}$ be the set of map functions present in our analytical model, where q numbers of Maps are executed in parallel. The intermediate *map()* output is shuffled among Reducers in the shuffle phase to achieve the data locality for better performance with respect to the faster execution. Let, $R = \{R_1, R_2, \ldots, R_r\}$ be the set of Reduces present in the MapReduce framework, where r *reduce()* functions exist. In general, number of Reduces are less than or equal to the number of Maps, i.e. $R(w) \leq Q(w)$. In our proposed data analysis model, analyzed data are moved to the output phase for storage and visualization purpose after the Reduce phase.

B. CORRELATION ANALYSIS USING MapReduce

The correlation analysis [32] is performed on the fine-grained processing of the data sets in the MapReduce framework. In our study, it is assumed that the health parameters are different in terms of type and number from one department to another. Let, $\Psi^{k}_{\psi_i} = \{\Psi^{k}_{\psi_1}, \Psi^{k}_{\psi_2}, \ldots, \Psi^{k}_{\psi_\delta}\}$, be the set of health parameters $\forall i \in DP$ and $\psi_1 \neq \psi_2 \neq \ldots \neq \psi_\delta$ in a hospital $k | \forall k \in H$. For example, $\Psi^{1}_{\psi_1}$ represents the parameter set 1 that belongs to the department 1 in the hospital 1.

In healthcare system, some health parameters are closely associated with each other with respect to the disease and its impact. Even, the relationships between the health variables are more complicated when a patient belongs to two or more departments in a hospital. Hence, we find the interrelationship between the patients having different health parameters and disease. In our proposed work, two types of analysis

are performed in correlation evaluation, i.e. *Intra-cluster* and *Inter-cluster* analysis as follows.

1) INTRA-CLUSTER CORRELATION ANALYSIS

In *Intra-cluster Correlation* [33], the patients within the same department are clustered based on their resemblance. Here, each individual is referred to as a patient p in a department δ, where $p \in P$ and $\delta \in DP$. Let ψ_δ be the number of health parameters associated in a department δ. Hence, the patient p has ψ_δ number of health parameters. However, all ψ_δ number of parameters are not equally responsible for a specific disease. Therefore, we find the correlation among those health parameters of a patient within the department DP_δ. Before correlation analysis, the health parameters (ψ_δ) are presented in the form of a health parameter matrix ($M^{Ph}_{\delta p}(w)$) as given in Eq. (10), where Ph represents the personal health of a patient p in a department δ within a window w.

$$M^{Ph}_{\delta p}(w) = {}^{P^k_{\delta p}}\begin{pmatrix} {}^{\Psi^k_{1\delta}} & {}^{\Psi^k_{2\delta}} & . & {}^{\Psi^k_{\psi\delta}} \\ (P^k_{\delta p}, \Psi^k_{1\delta}) & (P^k_{\delta p}, \Psi^k_{2\delta}) & . & (P^k_{\delta p}, \Psi^k_{\psi\delta}) \end{pmatrix} \quad (10)$$

Besides, within a department δ, p numbers of patients are available, where $p \in P$ with ψ_δ number of health parameters. The health parameters are stored in *intra-cluster* parameter matrix ($Ia^k_\delta(w)$) as given in Eq. (11).

$$Ia^k_\delta(w) = \begin{matrix} {}^{P^k_{\delta 1}} \\ {}^{P^k_{\delta 2}} \\ . \\ {}^{P^k_{\delta p}} \end{matrix} \begin{pmatrix} {}^{\Psi^k_{\psi\delta}} \\ M^{Ph}_{\delta 1}(w) \\ M^{Ph}_{\delta 2}(w) \\ . \\ M^{Ph}_{\delta p}(w) \end{pmatrix} \quad (11)$$

After simplification of Eq. (11), all health parameters related to the department δ can be represented in Eq. (12).

$$Ia^k_\delta(w) = \begin{matrix} {}^{P^k_{\delta 1}} \\ {}^{P^k_{\delta 2}} \\ . \\ {}^{P^k_{\delta p}} \end{matrix} \begin{pmatrix} {}^{\Psi^k_{1\delta}} & {}^{\Psi^k_{2\delta}} & . & {}^{\Psi^k_{\psi\delta}} \\ (P^k_{\delta 1}, \Psi^k_{1\delta}) & (P^k_{\delta 1}, \Psi^k_{2\delta}) & . & (P^k_{\delta 1}, \Psi^k_{\psi\delta}) \\ (P^k_{\delta 2}, \Psi^k_{1\delta}) & (P^k_{\delta 2}, \Psi^k_{2\delta}) & . & (P^k_{\delta 2}, \Psi^k_{\psi\delta}) \\ . & . & . & . \\ (P^k_{\delta p}, \Psi^k_{1\delta}) & (P^k_{\delta p}, \Psi^k_{2\delta}) & . & (P^k_{\delta p}, \Psi^k_{\psi\delta}) \end{pmatrix} \quad (12)$$

For example, *Cardiology* department (*Crd*) has multiple heart disease parameters with p number of patients. Now, the matrix $Ia^k_{Crd}(w)$ can be represented as given below.

$$Ia^k_{Crd}(w) = \begin{matrix} {}^{P^k_{Crd1}} \\ {}^{P^k_{Crd2}} \\ . \\ {}^{P^k_{Crdp}} \end{matrix} \begin{pmatrix} {}^{\Psi^k_{1\delta}=Age} & {}^{\Psi^k_{2\delta}=Sex} & . & {}^{\Psi^k_{\psi\delta}=thlh} \\ 63 & 1 & . & 150 \\ 67 & 1 & . & 108 \\ . & . & . & . \\ 41 & 0 & . & 172 \end{pmatrix}$$

In this work, our goal is to find the correlations ($\Gamma a^k_\delta(w)$) among different health parameters (ψ_δ) of patient p unlike the patients in a department δ within the window w. However, the health parameter correlation values are different from one patient to another within the same department due to the variance in the range of the parameters and severity of the disease. The correlation evaluation has different sub-steps such as column mean ($\rho a^k_\psi(w)$), variance ($\sigma a^k_\delta(w)$) and standard deviation ($SD^k_\delta(w)$). All those sub-steps are executed in various map phases and the *intra-cluster correlation* ($\Gamma a^k_\delta(w)$) is evaluated in the reduce phase. In the reduce phase, severity of the disease of a patient with respect to the correlated value of health parameters is checked and the high risk patients are clustered into a group (Ω_δ). The correlation factor $\Gamma a^k_\delta(w)$ lies always within the range [0, 1], where 1 and 0 are positive, and no correlation, respectively as given in Eq. (13).

$$\Gamma a^k_\delta(w) = \begin{cases} 1 & \textit{if positive correlation} \\ 0 & \textit{if no correlation} \end{cases} \quad (13)$$

Algorithm 1 Intra-cluster Correlation Evaluation (IaCE)

Input: χ : The size of each individual data partition.

Output: $\Gamma a^k_\delta(w)$: Intra-cluster correlation factor within window w.
$\Omega_\delta(w)$: Newly classified patient set within window w.

Notations:
p : # of patients in a hospital.
b : # of BANs in a hospital.
$\Phi^\delta_{Tot}(w)$: Total amount of health data collected in a healthcare cloud within window w.
ψ : # of health parameters associated with each patient and department.

1: Initialize $\Phi^\delta_{Tot}(w) = 0$;
2: e = p + b; // Total # of patients and BANs within the department δ.
3: **for** $\iota = 1$ **to** e **do**
4: **for** $\jmath = 1$ **to** ψ **do**
5: $\Phi^\delta_{Tot_{\iota\jmath}}(w)$ is calculated based on Eq. (5);
6: **end for**
7: **end for**
8: $U_p = \lfloor \frac{\Phi^\delta_{Tot}}{\chi} \rfloor$;
9: The steps from 10 to 17 are executed on U_p number of data blocks;
10: **for** $\iota = 0$ **to** $|U_p|$ **do**
11: **for** $\jmath = 0$ **to** ψ **do**
12: Intra-cluster matrix $Ia^k_\delta(w)[\iota][\jmath] = [P^k_\iota, \Psi^k_\jmath]$;
13: **end for**
14: Find the column mean $\rho a^k_\iota(w)$ based on the Eq. (14);
15: **end for**
16: Evaluate the variance ($\sigma a^k_\delta(w)$) based on Eq. (15);
17: Calculate standard deviation ($SD^k_\delta(w)$) based on Eq. (16);
18: Find Intra-cluster correlation ($\Gamma a^k_\delta(w)$) based on Eq. (17);
19: **if** $\Gamma a^k_\delta(w) \geq \Upsilon_a$ **then**
20: $\Omega_\delta(w) = \{P_\iota\}$;
21: **end if**
22: Return $\Gamma^\omega_a(w)$ and $\Omega_\delta(w)$;

The formal steps of *intra-cluster correlation* are described in Algorithm 1. Initially, the input and output parameters are

set in ***Intra-cluster Correlation Evaluation (IaCE)*** algorithm before its execution. According to the *IaCE* algorithm, the total collected data ($\Phi^{\delta}_{Tot}(w)$) is initialized. The total ($e = p + b$) number of patients and BANs are calculated within a department δ. Afterward, $\Phi^{\delta}_{Tot}(w)$ is calculated within the window w. In MapReduce model, each block size is fixed to be χ MB and is used as the input data set. Hence, U_p number of data blocks are generated and distributed among the active servers and all the statistical analysis are performed on those data blocks. The collected data are represented in an intra-cluster matrix ($Ia^k_{\delta}(w)$). The column mean ($\rho a^k_{\psi}(w)$) is calculated for all ψ number of parameters $\psi \in \Psi$ of all (e) number of patients and BANs present in the department δ, which is expressed in Eq. (14).

$$\rho a^k_{\psi}(w) = \frac{1}{\psi}\frac{1}{e}\sum_{\imath=1}^{\psi}\sum_{\jmath=1}^{e} Ia^k_{\delta}[\imath][\jmath] \tag{14}$$

Once the column mean is calculated in some $map_{\imath}()$ for each column in $Ia^k_{\delta}(w)$ and for $\imath \in Q$, the variance ($(\sigma a^k_{\delta}(w))^2$) and standard deviation ($SD^k_{\delta}(w)$) are evaluated within another $map_{\jmath}()$, where $\jmath \in Q$ and $\imath \neq \jmath$ within the window w for entire cluster matrix as expressed in Eq. (15).

$$(\sigma a^k_{\delta}(w))^2 = \frac{1}{e}\sum_{\jmath=1}^{\psi}\sum_{\imath=1}^{e}(P^k_{\imath}\psi_{\jmath} - \rho a^k_{\psi})^2 \tag{15}$$

After the variance, standard deviation ($SD^k_{\delta}(w)$) is calculated as given in Eq. (16).

$$SD^k_{\delta}(w) = \sqrt{(\sigma a^k_{\delta}(w))^2} \tag{16}$$

After $\rho a^k_{\psi}(w)$, $\sigma a^k_{\delta}(w)$ and $SD^k_{\delta}(w)$ are calculated in different map phases, the intermediate results are shuffled among the Reducers to achieve the data locality during execution. The intra-cluster correlation ($\Gamma a^k_{\delta}(w)$) is evaluated in the *reduce*() function as shown in Fig. 3 and is expressed in Eq. 17.

$$\Gamma a^k_{\delta}(w) = \frac{1}{e\psi}\sum_{\jmath=1}^{\psi}\sum_{\imath=1}^{e}\sum_{\aleph=1}^{e-1}\frac{(P^k_{\imath}\Psi_{\jmath} - \rho a^k_{\psi})(P^k_{\imath+\aleph}\Psi_{\jmath} - \rho a^k_{\psi})}{SD^k_{\delta}(w)} \tag{17}$$

Besides, the correlation factor ($\Gamma a^k_{\delta}(w)$) is checked by a threshold value (Υ_a) to know the severity of the parameters towards the disease. If the value of $\Gamma a^k_{\delta}(w)$ is higher than Υ_a, patients belonging to those highly influenced parameters are classified into a new set (Ω_{δ}). This newly classified set (Ω_{δ}) holds the high risk patients based on their health parameter correlated values for future health condition analysis and medications.

2) INTER-CLUSTER CORRELATION ANALYSIS

Inter-cluster correlation is used to find the similarity or dissimilarity between health parameters of different departments. For example, any heart patient $p \in P$ that has neurological disorder belongs to two different departments, i.e., *Cardiology* (Crd) department for *heart diseases* and *Neurology* (Neuro) department for the *neurological disorder*. According to two different departments, i.e. DP_{Crd} and DP_{Neuro}, the health parameters $\Psi^k_{\psi_{Crd}}$ and $\Psi^k_{\psi_{Neuro}}$ also varies with respect to the type as well as number. Hence, the *Inter-cluster* analysis is required to know the dependencies between $\Psi^k_{\psi_{Crd}}$ and $\Psi^k_{\psi_{Neuro}}$. Further, we also find the high risk patients and cluster them into a common set ($\Omega_{\imath\jmath}$) such as ($\Omega_{CrdNeuro}$) based on their correlated values. Let, $\psi_{\imath}$ and $\psi_{\jmath}$ be the number of health parameters associated with $\imath^{th}$ and $\jmath^{th}$ departments, respectively. Before *Inter-cluster correlation* analysis, the health parameters $\Psi^k_{\psi_{\imath}}$ and $\Psi^k_{\psi_{\jmath}}$ of $\imath^{th}$ and $\jmath^{th}$ departments are presented in *inter-cluster* parameter matrix ($Ie^k_{\psi_{\imath\jmath}}(w)$) within window w as given in Eq. (18).

$$Ie^k_{\psi_{\imath\jmath}}(w) = Ia^k_{\psi_{\imath}}(w) \cup Ia^k_{\psi_{\jmath}}(w) \tag{18}$$

The *Inter-cluster* correlation factor ($\Gamma e^k_{\imath\jmath}(t)$) ranges between [0, 1], where 1, and 0 are the positive and no correlation, respectively, as given in Eq. (19).

$$\Gamma e^k_{\imath\jmath}(t) = \begin{cases} +1 & \textit{if positive correlation} \\ 0 & \textit{if no correlation} \end{cases} \tag{19}$$

The ***Inter-cluster Correlation Evaluation (IeCE)*** is described in Algorithm 2. The input and output parameters are set initially in *IeCE* algorithm before the execution. Total $\Phi^{\imath\jmath}_{Tot}(w)$ amount of data are generated from $\imath^{th}$ and $\jmath^{th}$ departments within the window w. Let, $U^{\imath\jmath}_p$ be the number of data blocks generated from $\Phi^{\imath\jmath}_{Tot}(w)$ amount of data sets and is distributed among the active servers. Afterward, *Inter-cluster* matrix $Ie^k_{\psi_{\imath\jmath}}$ is set according to Eq. (18) in which $\psi_{\imath}$ and $\psi_{\jmath}$ amount of health parameters are present. Now, the column mean ($\rho e^k_{\psi_{\imath\jmath}}(w)$) is calculated for all parameters of both $\imath^{th}$ and $\jmath^{th}$ departments, which is expressed in Eq. (20).

$$\rho e^k_{\psi_{\imath\jmath}}(w) = \frac{1}{\psi_{\imath}\psi_{\jmath}}\sum_{i=1}^{\psi_{\imath}}\sum_{j=1}^{\psi_{\jmath}} Ie^k_{\psi_{\imath\jmath}}[i][j] \tag{20}$$

After calculation of the column mean, the variance ($(\sigma e^k_{\imath\jmath}(w))^2$) is computed within different map functions in window w, which can be given in Eq. (21).

$$(\sigma a^k_{\imath\jmath}(w))^2 = \frac{1}{\psi_{\imath}\psi_{\jmath}}\sum_{i=1}^{\psi_{\imath}}\sum_{j=1}^{\psi_{\jmath}}(Ie^k_{\psi_{\imath\jmath}}[i][j] - \rho e^k_{\psi_{\imath\jmath}})^2 \tag{21}$$

Hence, the standard deviation ($SD^k_{\imath\jmath}(w)$) is measured for $\imath^{th}$ and $\jmath^{th}$ departments and is expressed in Eq. (22).

$$SD^k_{\imath\jmath}(w) = \sqrt{(\sigma a^k_{\imath\jmath}(w))^2} \tag{22}$$

Thereafter, the inter-cluster correlation ($\Gamma e^k_{\imath\jmath}(w)$) is evaluated in the *reduce*() function after $\rho e^k_{\psi_{\imath\jmath}}(w)$, $(\sigma e^k_{\imath\jmath}(w))^2$ and $SD^k_{\imath\jmath}(w)$ calculations are finished in different map phases. The intermediate results are shuffled among the Reducers for

Algorithm 2 Inter-Cluster Correlation Evaluation (IeCE)

Input: χ : The size of each individual data partition.

Output: $\Gamma e_{\delta}^{k}(w)$: Inter-cluster correlation factor at time t.
$\Omega_{\iota J}(w)$: Newly classified patient set at time t.

Notations:
$\Phi_{Tot}^{\iota J}(w)$: Total amount of health data collected from ι^{th} and J^{th} department within window w.
ψ_{ι} : # of health parameters associated with ι^{th} department.
ψ_{J} : # of health parameters associated with J^{th} department.

1: Initialize $\Phi_{Tot}^{\iota J}(w) = 0$;
2: **for** $\iota = 1$ **to** ψ_{ι} **do**
3: **for** $J = 1$ **to** ψ_{J} **do**
4: $\Phi_{Tot}^{\iota J}(w)$ amount of data are collected;
5: **end for**
6: **end for**
7: $U_{p}^{\iota J} = \lfloor \frac{\Phi_{Tot}^{\iota J}}{\chi} \rfloor$;
8: The steps from 9 to 21 are executed on $U_{p}^{\iota J}$ number of data blocks;
9: **for** $\iota = 0$ **to** ψ_{ι} **do**
10: **for** $J = 0$ **to** ψ_{J} **do**
11: Intra-cluster matrix $Ie_{\psi_{\iota J}}^{k}(w)[\iota][J] = [\Psi_{\iota}^{k}, \Psi_{J}^{k}]$;
12: **end for**
13: Find the column mean $\rho e_{\psi_{\iota J}}^{k}(w)$ based on the Eq. (20);
14: **end for**
15: **for** $\iota = 0$ **to** P **do**
16: **for** $J = 0$ **to** Ψ **do**
17: Inter-cluster matrix $I_{e}^{\omega}(w)[\iota][J] = P_{\iota}\Psi_{J}$;
18: **end for**
19: Find the column mean $\rho_{\psi}^{\omega}(w)[\iota] = \frac{1}{P}(P_{\iota})(\Psi_{\iota})$;
20: **end for**
21: Compute variance $((\sigma e_{\iota J}^{k}(w))^{2})$ based on Eq. (21);
22: Calculate standard deviation $(SD_{\iota J}^{k}(w))$ based on Eq. (22);
23: Find *inter-cluster correlation* $(\Gamma e_{\iota J}^{k}(w))$ based on Eq. (23);
24: **if** $\Gamma e_{\iota J}^{k}(w) \geq \Upsilon_{e}$ **then**
25: $\Omega_{\iota J}(w) = \{P_{\iota}\}||\{P_{J}\}$;
26: **end if**
27: Return $\Gamma e_{\iota J}^{k}(w)$ and $\Omega_{\iota J}(w)$;

data locality as shown in Fig. 3, where $r_{\iota} \in R$. The inter-cluster correlation $(\Gamma e_{\iota J}^{k}(w))$ is expressed in Eq. 23.

$$\Gamma e_{\iota J}^{k}(w) = \frac{1}{\psi_{\iota}\psi_{J}} \times \sum_{i=1}^{\psi_{\iota}} \sum_{j=1}^{\psi_{J}} \sum_{\aleph=1}^{e-1} \frac{(\Psi_{i}^{k}\Psi_{j}^{k} - \rho e_{\psi_{ij}}^{k})(\Psi_{i+\aleph}^{k}\Psi_{j+\aleph}^{k} - \rho e_{\psi_{ij}}^{k})}{SD_{\iota J}^{k}(t)} \tag{23}$$

In addition to the inter-cluster correlation analysis, the highly influenced health parameters are identified and grouped together based on the correlation values. A threshold

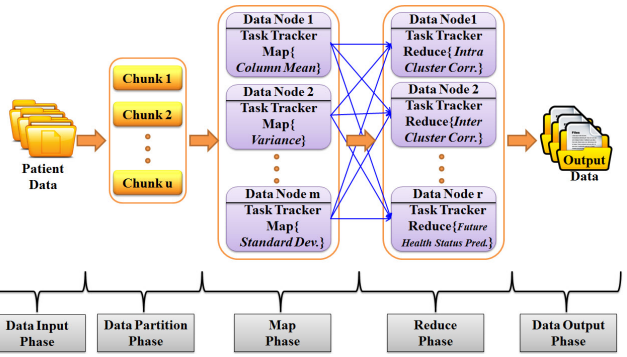

FIGURE 3. MapReduce model for healthcare data analysis and processing.

(Υ_{e}) is set as the high risk level and is compared with the correlation values. If the value of $\Gamma e_{\iota J}^{k}(w)$ is higher than the Υ_{e}, the respective patients with those health parameters are classified into a new group $(\Omega_{\iota J})$. This newly classified group is considered as the high risk patients who need proper medications and precautions.

VI. FUTURE HEALTH PREDICTION MODEL

In this section, we predict the future health status of the patients based on their current health parameters (Ψ). Note that the patients are grouped together in a set $(\Omega_{\iota J})$ in a particular department δ based on their correlated values $(\Gamma a_{\delta}^{k}(w))$. During the diagnosis, many questions may be asked by the doctors related to the past history of the patients to know a patient's current health condition. The doctors also inspect the hidden symptoms related to the diseases of the patients. However, the symptoms may vary from patient to patient with different severity. By understanding the above system, a Hidden Markov Model (HMM) [34], [35] is formulated and a Viterbi algorithm [36] is used to know the most likely sequence of the hidden states.

Let, $S = \{S_1, S_2, \ldots, S_K\}$ and $O = \{O_1, O_2, \ldots, O_L\}$ be the state and observe space set in HMM, respectively. The *Flu* patients' future health prediction is considered as the application of our proposed prediction model. Let, the i^{th} *Flu* patient (P_i) be treated by the j^{th} doctor (D_j) having ψ health parameters within window w_0 as the initial case. A sequence of hidden states $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_w\}$ such as a runny nose (Rn), Sneezing (Sn), Strep Throat (St), i.e. $\gamma = \{\gamma_{Rn}, \gamma_{Sn}$ and $\gamma_{St}\}$ are found by the doctor at different time instances during observations. However, it is assumed that each observation is associated with different probabilities to make the environment more realistic. Without losing generality, the initial probability (Π) is considered, where $\Pi = \{\Pi_1, \Pi_2, \ldots, \Pi_K\}$. Let, $\alpha_{\iota J}$ be the transition probability from state ι to state J. An emission probability $\beta_{\iota J}$ is defined to estimate how likely the patient feels during observation O_J on each arrival time, which affects the state S_t.

The Max-product algorithm [37] known as Viterbi algorithm is used to find the posterior probability (future health

status). Let, Λ be the most probable state sequence of observations for future prediction, where $\Lambda = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_w\}$. The exact path of the most probable state sequence is stored in η as a variable, where $\eta = \{\eta_1, \eta_2, \ldots, \eta_w\}$. In case of healthcare data analysis, future health condition prediction (state transition) is hard to compute accurately with standard HMM due to diversified nature of the data sets. Hence, the Intra-cluster correlation ($\Gamma a^k_\delta(w)$) coefficient is added as a new parameter to the standard HMM in order to improve the efficiency of the prediction model. By applying the recurrences, following equations can be derived.

$$\Lambda_{11}(1) = \beta(\gamma_1|S_1) * \Gamma a^k_{\delta_\ell}(w) * \Pi_1, \text{ for } t,\ K,\ \text{and } L = 1. \tag{24}$$

For L observations and K states within window w, we get

$$\left.\begin{aligned} \Lambda_{LK}(w) &= \beta(\gamma_L \mid S_K) * \alpha_{\iota K} * \Gamma a^k_{\delta_\ell}(w) * \Lambda_{L-1K-1}(w-1) \\ \eta_{LK}(w) &= \arg\max_{\iota \in S}\{\Lambda_{LK}(w)\} \end{aligned}\right\} \tag{25}$$

Algorithm 3 Future Health Condition Prediction (FHCP)

Input: Π : Initial probability where $\Pi = \{\Pi_1, \Pi_2, \ldots, \Pi_\ell\}$.
$\alpha_{\iota\jmath}$: Transition probability from ι^{th} state to $\jmath^{th}$ state.
$\beta_{\iota\jmath}$: Emission probability of ι^{th} state and $\jmath^{th}$ observation.
$\Gamma a^k_{\delta_\ell}(w)$: Intra-cluster Correlation where ℓ is the # of patients available within the risk set Ω_δ.

Output: Λ : Most likely hidden state sequence where $\Lambda = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_w\}$.
η : Store the most probable path where $\eta = \{\eta_1, \eta_2, \ldots, \eta_w\}$.

Notations:
S : States where S $= \{S_1, S_2, \ldots, S_\ell\}$.
O : Observation where O $= \{O_1, O_2, \ldots, O_L\}$.
γ : Sequence of hidden state observations, where $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_w\}$.

1: Initialize
$\Lambda = 0$ and $\eta = 0$
2: Find the Intra-cluster correlation($\Gamma a^k_\delta(w)[\ell]$) = *IaCE()*;
3: **for** $\iota = 0$ **to** ℓ **do**
4: Calculate $\Lambda_{\iota 1} = \Pi_\iota * \Gamma a^k_{\delta_\iota}(w) * \beta_{\iota 1}$; (base cases)
5: Calculate $\eta_{\iota 1} = \iota$;
6: **end for**
7: **for** $\jmath = 0$ **to** w **do**
8: **for** $\iota = 0$ **to** ℓ **do**
9: $\Lambda_{\iota\jmath} = \max_\ell(\Lambda_{\iota\jmath-1} * \alpha_{\ell\jmath} * \Gamma a^k_{\delta_\iota}(w) * \beta_{\jmath\gamma_t})$; # $w > 1$
10: Calculate $\eta_{\iota\jmath} = \arg\max_\ell(\Lambda_{\iota\jmath})$;
11: **end for**
12: return Λ *and* η;
13: **end for**

As given in Algorithm 3, patients having *Flu* are diagnosed and the future health status is checked as the application of Future Health Condition Prediction (FHCP) algorithm. In a specific example, suppose a patient $p \in P$ has two states, i.e. healthy or flu, where $S = \{S_{Healthy}, S_{Flu}\}$. The health status is calculated based on some observations (O) such as in day 1, day 2 etc, and hidden symptoms (γ) such as a runny nose (Rn), Sneezing (Sn), Strep Throat (St), i.e. $\gamma = \{\gamma_{Rn}, \gamma_{Sn}$ and $\gamma_{St}\}$. All input and output health notations of patients are initialized in *FHCP* algorithm. The initial probability (Π) is set for the state change from starting point to the health state, where $\Pi = \{\Pi_{Healthy} : 0.5,\ \Pi_{Flu} : 0.5\}$. Further, the intra-cluster correlated ($\Gamma a^k_\delta(w)$) values are calculated to find the similar group of patients. Hence, the future state is estimated by considering only those correlated patients instead of all within the department. The correlation value ($\Gamma a^k_\delta(w)$) is set to be 0.5 in our evaluation. However, this value may vary with respect to patients and time as well. The state change probability or transition probability ($\alpha_{Healthy,Flu}$) is set for the state change from healthy to flu and viceversa.

$$\alpha_{Healthy,Flu} = \begin{array}{c} \\ Healthy \\ Flu \end{array} \begin{array}{c} \begin{array}{cc} Healthy & Flu \end{array} \\ \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix} \end{array}$$

Similarly, the emission probability $\beta_{Healthy,Flu}$ is set according to the hidden symptoms for two states.

$$\beta_{Healthy,Flu} = \begin{array}{c} \\ Healthy \\ Flu \end{array} \begin{array}{c} \begin{array}{cccc} Rn & Sn & St & Normal \end{array} \\ \begin{pmatrix} 0.3 & 0.1 & 0.1 & 0.5 \\ 0.1 & 0.2 & 0.6 & 0.1 \end{pmatrix} \end{array}$$

In the first day of observation, let health condition of a patient be good as observed by the doctor, i.e. normal: 0.5 and the most probable state (Λ) is calculated for the initial step. $\Lambda^1_{Start,Healthy} = \Pi_{Healthy}$: 0.5 * $\beta_{Healthy,Normal}$: 0.5 * $\Gamma a^k_\delta[\iota](t)$: 0.5. Hence, $\Lambda^1_{Start,Healthy} = 0.125$. Similarly, $\Lambda^1_{Start,Flu} = \Pi_{Flu}$: 0.5 * $\beta_{Flu,Normal}$: 0.1 * $\Gamma a^k_\delta[\iota](t)$: 0.5. Hence, $\Lambda^1_{Start,Flu} = 0.025$. Then, the exact path of the most probable state sequence is stored in η, which is selected as the max{Λ} for the initial step of the algorithm. Therefore, $\eta_1 = \max\{\Lambda^1_{Start,Healthy} : 0.125,\ \Lambda^1_{Start,Flu} : 0.025\}$. Finally, $\eta_1 = \{S_{Healthy} : 0.125\}$.

However, in the next day of the visit, patient $p \in P$ is having the runny nose (Rn). Again, the probable state (Λ), and path (η) are calculated recursively. By considering the day one probabilities, i.e. $\Lambda^1_{Start,Healthy}$ and $\eta_1 = S_{Healthy}$, we can get $\Lambda^2_{Healthy,Healthy} = \Lambda^1_{Start,Healthy}$: 0.125 * $\alpha_{Healthy,Healthy}$: 0.6 * $\beta_{Healthy,Rn}$: 0.3 * $\Gamma a^k_\delta[\iota](t)$: 0.5. Hence, $\Lambda^2_{Healthy,Healthy} = 0.01125$. Similarly, $\Lambda^2_{Healthy,Flu} = \Lambda_{Start,Flu}$: 0.025 * $\alpha_{Healthy,Flu}$: 0.4 * $\beta_{Flu,Rn}$: 0.1 * $\Gamma a^k_\delta[\iota](t)$: 0.5. Therefore, $\Lambda^2_{Start,Flu} = 0.0005$. Now, η is updated by selecting the value of max{Λ} and the most probable state is stored. Therefore, $\eta_2 = \max\{\Lambda^2_{Healthy,Healthy} : 0.01125,\ \Lambda^2_{Healthy,Flu} : 0.0005\}$.

Let, the patient is having Strep Throat (*St*) with runny nose (*Rn*) on the third day. Here, the previous day symptoms are also considered by multiplying the previous day observed probability with the current day observed probability. For example, *Rn* probability 0.3 in day 3 is multiplied with the

probability of St, which is 0.1 for $\Lambda^3_{Healthy,Healthy}$ calculation. Hence, $\Lambda^3_{Healthy,Healthy} = \Lambda^2_{Healthy,Healthy} : 0.01125 * \alpha_{Healthy,Healthy}$: 0.6 * $\beta_{Healthy,Rn}$: 0.3 * $\beta_{Healthy,St}$: 0.1 * $\Gamma a^k_\delta[\iota](t)$: 0.5. As a result, $\Lambda^3_{Healthy,Healthy} = 0.00010125$. Similarly, $\Lambda^3_{Healthy,Flu} = 0.000006$. By considering the highest vale, $\eta_3 = \Lambda^3_{Healthy,Healthy} = 0.00010125$. Let the patient has sneezing (Sn) in the fourth day of the observation. By considering these hidden symptoms, the most probable state is calculated. Hence, $\Lambda^4_{Healthy,Healthy} = 2.025e\text{-}7$ and $\Lambda^4_{Healthy,Flu} = 2.16e\text{-}7$. If we take logarithm of both states, $\Lambda^4_{Healthy,Healthy} = \log(2.025e\text{-}7) = -6.693$ and $\Lambda^4_{Healthy,Flu} = \log(2.16e\text{-}7) = -6.665$. According to the rule, the highest value is taken as the state of the day, which is $\eta_4 = \Lambda^4_{Healthy,Flue} = -6.665$. Hence, it is concluded that the patient is having flu after the fourth day. By using backtracking, the most probable state paths are found from the predicted to the starting day. The above calculations are only for one patient. However, to handle a large number of patients, we can run *FHCP algorithm* separately in different *map* phases for different patients.

VII. PERFORMANCE EVALUATION

In this section, evaluations of the proposed algorithms are carried out by using CloudSim 3.0 [38] with Java Eclipse Integrated Development Environment (IDE). All simulations are performed on the Intel core i7 3.4GHz systems to process the patient data in cloud MapReduce [39]. For our simulation, healthcare patient data are taken from publicly available machine learning repository in the center for machine learning and intelligent systems [40]. In our simulation, cardiac (heart diseases) patients data are considered for *IaCE* and *IeCE* algorithm, whereas flu patients data are used for *FHCP* algorithm execution. All the data are collected from Cleveland and Hungarian clinic [40] available for public use. During our analysis, 14 number of heart disease attributes are used including the structured, semi-structured and unstructured data. The step-by-step procedure of CloudSim simulator is shown in Fig. 4.

In CloudSim simulator, the packages and library files are initialized in the first step, which are going to be used in the simulation process. By using *CloudSim.init (num_user, calendar, trace_flag)* function, the library files are initialized, where *num_user* represents the number of cloud users, *calendar* holds the current date and *trace_flag* is used to print the events as shown in step 1 of Fig. 4.

Data centers are the backbone of the cloud environment, where each data center is comprised of multiple hosts. As shown in step 2 of Fig. 4, *Datacenter datacenter0 = createDatacenter ("Datacenter_0")* is used to create the data center. Resource provisioning to virtual machines is the basic objective of the host and the detail list of parameters for the data center host is shown in Table 1. It is assumed that the data centers are networked and geographically distributed. In step 3 of Fig. 4, the broker is created by *DatacenterBroker broker = createBroker()* method and acts as a user in the data

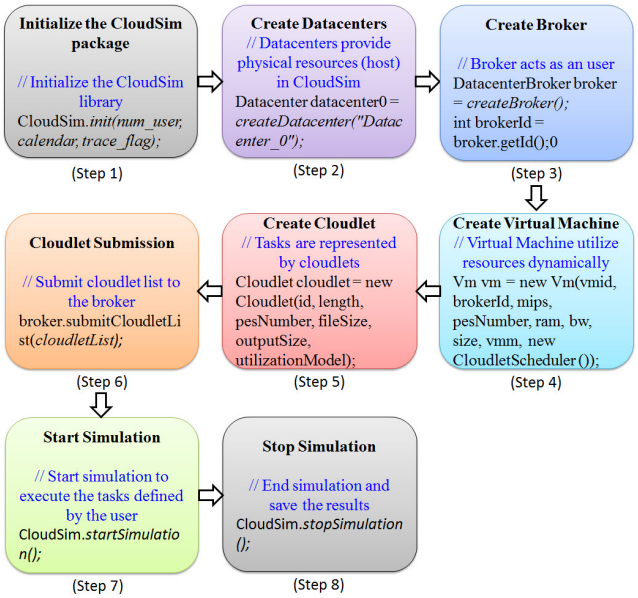

FIGURE 4. Simulation steps of CloudSim simulator.

TABLE 1. Parameter list of data centers

Parameter list of data centers	
Parameters	**Values**
System Architecture	x86
Operating System	Linux
MIPS	1000
VMM	Xen
PEs(CPU)	1
Primary Memory (RAM)	2048 Mb
Secondary Storage	1000000 Mb
Bandwidth	10000 Mbps
VmScheduler	TimeShared

TABLE 2. Parameter list of virtual machines

Parameter List of Virtual Machines	
Parameters	**Values**
MIPS	1000
Size	10000
VMM	Xen
PEs(CPU/Cores)	1
Primary Memory (RAM)	512 Mb
Bandwidth	1000 Mbps
VmScheduler	TimeShared

center. The broker is responsible for assigning the Virtual Machines (VMs) to the hosts (Physical Machines) and also sets cloudlets (tasks) to the VMs. Each VM is created in step 4 of Fig. 4 to handle the service tasks by sharing the physical machines with respect to time and resources. The parameters of the VMs are defined in Table 2 and are set as follows, *Vm vm = new Vm (vmid, brokerId, mips, pesNumber, ram, bw, size, vmm, new CloudletSchedulerTimeShared())*.

The cloudlets (or tasks) are created in step 5 as shown in Fig. 4. The basic parameters of cloudlets are shown in Table 3 and are set in CloudSim as follows: *Cloudlet coudlet = new Cloudlet (id, length, pesNumber, fileSize, outputSize, utilizationModel, utilizationModel, utilizationModel)*.

TABLE 3. Parameter list of cloudlets

Parameter List of Cloudlets(Tasks)	
Parameters	**Values**
Length	4000
Input File Size	300 Mb
Output File Size	300 Mb
Utilization Model	Full utilization

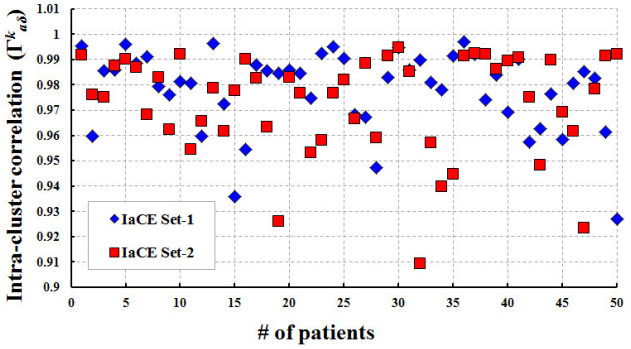

FIGURE 5. Intra-cluster correlation evaluation values.

Here, our own Map ($q \in Q$) and Reduce ($r \in R$) functions are defined for execution, where each task has either one Map or Reduce function represented as a cloudlet. However, the Map and Reduce functions can be defined externally and import the outputs in CloudSim IDE. In the next step, i.e., step 6 of Fig. 4, the cloudlet is submitted to the broker present in the data center by calling *broker.submitCloudletList (cloudletList)*. Hence, the simulation will start on the next step (step 7) of Fig. 4 by calling *CloudSim.startSimulation()*. Finally, the simulation is stopped in step 8 of Fig. 4 and the results are stored by calling *CloudSim.stopSimulation()*.

A. SIMULATION RESULTS

The pre-processing step is performed to normalize the raw data for execution in some reserved *map*() $\in Q$. As shown in Fig. 5, the intra-cluster correlated (Γa^k_δ) values are plotted for two sets of patients by executing the *IaCE* algorithm. Instead of all the patients, 50 number of patients are considered in each set to observe the data more clearly. The plot shows that almost all the patients correlation values are greater than 0.95, which leads to $+ve$ correlation exist between two sets in one cluster. Therefore, the healthcare attributes of set 1 are highly correlated with the attributes of set 2. From the figure, it seems that our proposed algorithm is efficient for correlation analysis of the heart disease patients.

In Fig. 6, the *inter-cluster correlated* (Γe^k_{IJ}) values are plotted for two sets of patients belong to two different clusters, i.e. Cleveland and Hungarian by executing *IeCE_(Cleveland - Hungarian)* algorithm on different sets of patients other than the sets of patients used in Fig. 5. In this graph, *IaCE_Cleveland* and *IaCE_Hungarian* values are also plotted to compare with *IeCE_(Cleveland-Hungarian)*. It is observed that some correlation values are less than 0.6

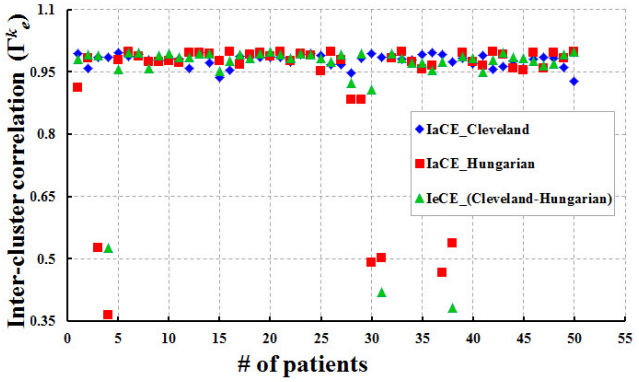

FIGURE 6. Inter-cluster correlation evaluation values.

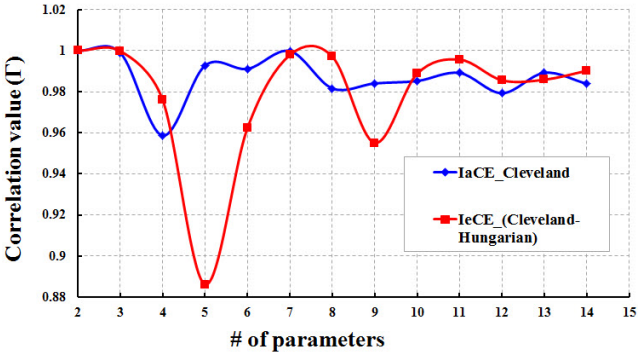

FIGURE 7. Correlation variances with # of parameters.

and most of them are greater than 0.95. The values less than 0.6 are treated as the less correlated values, where greater than 0.9 values are treated as highly correlated patients. This correlation analysis could be extended to a classification analysis based on the correlation values.

The number of attributes explained in Fig. 7 plays an important role in the correlation analysis. According to the simulation result, initially both *intra-cluster correlation*, and *inter-cluster correlation* analysis are varied due to less number of parameters. However, this variation is minimized when the number of parameters are increased. *IeCE_(Cleveland - Hungarian)* correlation value is more drifted than *IaCE_(Cleveland)* correlation value as *IeCE_(Cleveland - Hungarian)* is executed with two different clusters and *IaCE_(Cleveland)* is carried out within a single cluster. It is clearly observed that the correlation value is more stable with more number of parameters.

In Fig. 8, the step by step execution of *FHCP* algorithm is exhibited, where *Viterbi path* is estimated by considering the current health condition of the patients. In the initial stage of *FHCP*, the health states and patient's observations are determined. In this scenario, the starting probability is set to be 0.5 for both *Healthy* and *Flu* state by giving equal opportunity. Similarly, the correlation value is set to be 0.5 at initial time instance. The initial state probability is calculated by multiplying the initial probability, correlation factor with the state transition probability and found that *Healthy* state

```
CloudSim Execution Started..!!
Initialising...
Starting CloudSim version 3.0
Datacenter_0 is starting...
Datacenter_1 is starting...
Broker is starting...
Entities started.
0.0: Broker: Cloud Resource List received with 2 resource(s)
0.0: Broker: Trying to Create VM #0 in Datacenter_0
0.0: Broker: Trying to Create VM #1 in Datacenter_0
[VmScheduler.vmCreate] Allocation of VM #1 to Host #0 failed by MIPS
0.1: Broker: VM #0 has been created in Datacenter #2, Host #0
0.1: Broker: Creation of VM #1 failed in Datacenter #2
0.1: Broker: Trying to Create VM #1 in Datacenter_1
0.2: Broker: VM #1 has been created in Datacenter #3, Host #0
0.2: Broker: Sending cloudlet 0 to VM #0
0.2: Broker: Sending cloudlet 1 to VM #1
160.2: Broker: Cloudlet 0 received
160.2: Broker: Cloudlet 1 received
160.2: Broker: All Cloudlets executed.

Future Health State Prediction (FHSP) starts execution..!!
States: {Healthy, Flu}
Observations: {Normal, Sneezing, Strep Throat, Running Nose}
Start probability: {Healthy: 0.5, Flu: 0.5}
Transition probability:
 Healthy: {  Healthy: 0.6,   Flu: 0.4, }
 Flu: {  Healthy: 0.4  Flu: 0.6}

Emission probability:
 Healthy: {  Normal: 0.5,   Sneezing: 0.1,   Strep Throat: 0.1,   Running Nose: 0.3, }
 Flu: {  Normal: 0.1  Sneezing: 0.1  Strep Throat: 0.6  Running Nose: 0.2}

Viterbi path: [Day1:Healthy -> Day2:Healthy -> Day3:Flu -> Day4:Flu -> Day5:Flu].

Future Health State Prediction (FHSP) ends execution..!!

160.2: Broker: Destroying VM #0
160.2: Broker: Destroying VM #1
Broker is shutting down...
Simulation: No more future events
CloudInformationService: Notify all CloudSim entities for shutting down.
Datacenter_0 is shutting down...
Datacenter_1 is shutting down...
Broker is shutting down...
Simulation completed.
Simulation completed.

========== OUTPUT ==========
Cloudlet ID    STATUS    Data center ID    VM ID    Time    Start Time    Finish Time
    0        SUCCESS        2            0        160        0.2        160.2
    1        SUCCESS        3            1        160        0.2        160.2
CloudSim Execution finished..!!
```

FIGURE 8. Viterbi path estimation for future health prediction.

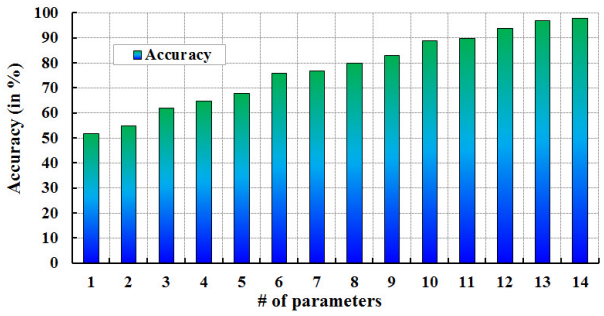

FIGURE 9. Accuracy of Viterbi path prediction.

has the higher probability than the *Flu* one. Therefore, in Day 1, the patient is felt healthy. Similarly, on the next day, i.e. Day 2 also the health condition of the patient is *Healthy*. However, the patient suffers from *Flu* in Day 3 as the hidden symptom (*Strep Throat: 0.6*) has higher probability, which affects the change of state. By continuing *FHCP* algorithm, the patient condition is estimated to take necessary precaution and preventions.

It is clearly noticed in Fig. 9 that the accuracy of prediction is also increased with increase in the number of attributes. When only one attribute is considered, the accuracy is touched around 51%. However, it is increased up to 58% with two attributes. The accuracy is about 80%, when the numbers of attributes are increased to 8 and eventually we got the maximum accuracy of 98%, when the numbers of attributes are increased to 14.

Processing time is an important factor to observe the efficiency of the algorithms, which is shown in Fig. 10. Here, the processing time is defined as the summation of task

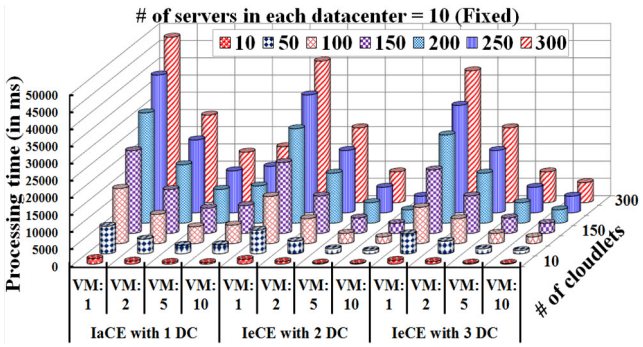

FIGURE 10. Processing time.

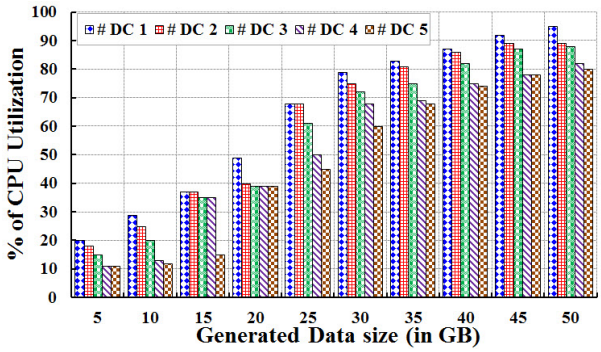

FIGURE 11. CPU utilization with increase in data size.

scheduling time with data transfer time from different data centers to achieve the data locality and execution time. It is observed that the processing time is reduced by using multiple virtual machines in different data centers. The processing time of *IaCE* algorithm is longer than the *IeCE* algorithm as the tasks within one data center wait for the execution until the running tasks are finished. The *IeCE* algorithm with 3 data centers is faster than the *IeCE* algorithm with 2 data centers due to parallel processing of the map tasks. Here, the main advantage of using cloud platform is to reduce the processing time. For example, let us consider one VM in a standalone system without any cloud platform. The processing time is longer for *IaCE* algorithm with this configuration, where the number of cloudlets are more. However, if we increase the number of VMs, obviously the processing time is reduced drastically. Similar trend can be observed for *IeCE* algorithm with different numbers of VMs and configuration. However, always increase in number of data centers does not enhance the processing time as the data are sparse and require more time for execution. Therefore, the processing time is directly proportional to the execution and data transfer time.

The CPU utilization during processing of huge patient data in the data centers is shown in Fig. 11. In our simulation, the input data sizes are set in gigabytes that range from 5GB to 50GB. From the simulation result, higher CPU utilization is observed with an increase of the amount of data size coming to the data centers, which continues until it reaches at the processing threshold. If the tasks have a hard deadline, new data centers are added in the execution process to balance

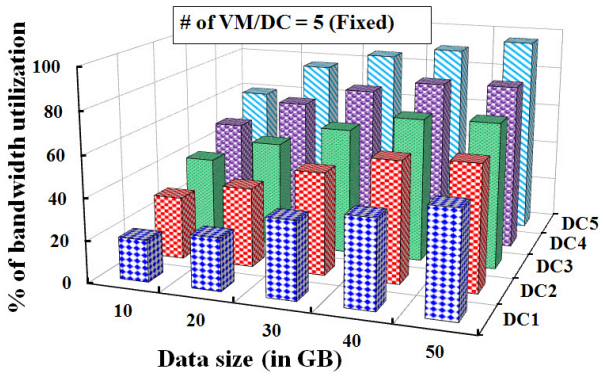

FIGURE 12. Bandwidth utilization with increase in data size.

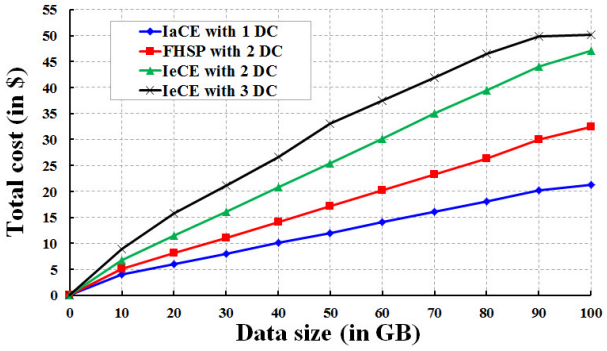

FIGURE 13. Total cost estimation.

the under and over utilization. The main objective here is to maximize the resource utilization without compromising the processing deadline though we need to check for the under and over utilization to minimize the cost.

In Fig. 12, the bandwidth utilization in the cloud service is displayed. Here, 5 VMs are created and are fixed for each data centers. It is observed that the percentage of bandwidth utilization depends on the data size and geographically distributed data centers associated with the cloud services. For instance, initially, the bandwidth utilization of DC5 was 59% for 10GB of data. But, the utilization eventually boosts up to 96% with 50GB of data. However, the bandwidth directly depends on the network traffic and time variant in nature, which affects the bandwidth cost estimation.

From revenue point of view, cost is another major factor, which cannot be ignored and is shown in Fig. 13. In the simulation, bandwidth cost, storage cost, computation cost and data migration cost are taken into account. For cost calculation, Amazon Web Service (AWS) pricing model is taken as the reference and through simulation, the cost per processing data in GB came out to be 0.5USD approximately. The *IaCE* cost is less than the *FHCP* and *IeCE*. Most of the cost is varied due to data transfer among different data centers, which is directly proportional to the total cost. However, the growth rate of cost per GB does not follow the same trend in the growth rate of arrival data. For instance, the total cost is increased linearly for processing the data up to 90GB though it becomes steady between 90GB to 100GB.

VIII. CONCLUSION

In this paper, a probabilistic data acquisition method is designed for the cloud based healthcare system. Besides, *IaCE* and *IeCE* algorithms are designed for the intra and inter cluster correlation analysis of the healthcare big data. An *FHCP* algorithm is designed to predict the future health condition of the patients based on their current health status with the accuracy of 98%. In addition, cloud-based MapReduce model is used as the processing framework for our big data analysis. It is observed that our protocol can be used for various applications related to healthcare and patient monitoring such as heart disease prediction or cancer severity classification. Our future work is to implement the proposed data analytic model in the real healthcare domain to analyze the data in real-time data analytic platform such as SPARK.

REFERENCES

[1] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The Internet of Things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, "Big data for health," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1193–1208, Jul. 2015.

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Inf. Sci. Syst.*, vol. 2, no. 1, pp. 1–10, 2014.

[4] (Apr. 2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf

[5] K. Lin, F. Xia, W. Wang, D. Tian, and J. Song, "System design for big data application in emotion-aware healthcare," *IEEE Access*, vol. 4, pp. 6901–6909, 2016.

[6] L. A. Tawalbeh, R. Mehmood, E. Benkhlifa, and H. Song, "Mobile cloud computing model and big data analysis for healthcare applications," *IEEE Access*, vol. 4, pp. 6171–6180, 2016.

[7] C. K. Dehury and P. K. Sahoo, "Design and implementation of a novel service management framework for iot devices in cloud," *J. Syst. Softw.*, vol. 119, pp. 149–161, Sep. 2016.

[8] Z. Yu *et al.*, "Incremental semi-supervised clustering ensemble for high dimensional data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 701–714, Mar. 2016.

[9] M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.

[10] S. Rallapalli, R. R. Gondkar, and U. P. K. Ketavarapu, "Impact of processing and analyzing healthcare big data on cloud computing environment by implementing hadoop cluster," *Procedia Comput. Sci.*, vol. 85, pp. 16–22, May 2016.

[11] S. Wang, X. Chang, X. Li, G. Long, L. Yao, and Q. Z. Sheng, "Diagnosis code assignment using sparsity-based disease correlation embedding," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3191–3202, Dec. 2016.

[12] V. Tresp, J. M. Overhage, M. Bundschus, S. Rabizadeh, P. A. Fasching, and S. Yu, "Going digital: A survey on digitalization and large-scale data analytics in healthcare," *Proc. IEEE*, vol. 104, no. 11, pp. 2180–2206, Nov. 2016.

[13] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," *Big Data Res.*, vol. 2, no. 1, pp. 2–11, 2015.

[14] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008.

[15] C.-H. Wu and Y.-C. Tseng, "Data compression by temporal and spatial correlations in a body-area sensor network: A case study in Pilates motion recognition," *IEEE Trans. Mobile Comput.*, vol. 10, no. 10, pp. 1459–1472, Oct. 2011.

[16] R. A. Taylor *et al.*, "Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data–driven, machine learning approach," *Acad. Emerg. Med.*, vol. 3, no. 23, pp. 269–278, Mar. 2016.

[17] M. Patel and J. Wang, "Applications, challenges, and prospective in emerging body area networking technologies," *IEEE Wireless Commun.*, vol. 17, no. 1, pp. 80–88, Feb. 2010.
[18] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.
[19] F. Zhang, J. Cao, S. U. Khan, K. Li, and K. Hwang, "A task-level adaptive mapreduce framework for real-time streaming data in healthcare applications," *Future Generat. Comput. Syst.*, vols. 43–44, pp. 149–160, Feb. 2015.
[20] Y. Zhang, S. Chen, Q. Wang, and G. Yu, "i^2 MapReduce: Incremental MapReduce for mining evolving big data," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 7, pp. 1906–1919, Jul. 2015.
[21] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua, "Disease inference from health-related questions via sparse deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 8, pp. 2107–2119, Aug. 2015.
[22] M. Barkhordari and M. Niamanesh, "ScaDiPaSi: An effective scalable and distributable mapreduce-based method to find patient similarity on huge healthcare networks," *Big Data Res.*, vol. 2, no. 1, pp. 19–27, 2015.
[23] C.-H. Weng, T. C.-K. Huang, and R.-P. Han, "Disease prediction with different types of neural network classifiers," *Telematics Inform.*, vol. 33, no. 2, pp. 277–292, 2016.
[24] S. Gopakumar, T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh, "Stabilizing high-dimensional prediction models using feature graphs," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 1044–1052, May 2015.
[25] H. Li, X. Li, M. Ramanathan, and A. Zhang, "Prediction and informative risk factor selection of bone diseases," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 1, pp. 79–91, Jan./Feb. 2015.
[26] J. Henriques *et al.*, "Prediction of heart failure decompensation events by trend analysis of telemonitoring data," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1757–1769, Sep. 2015.
[27] P. Gope and T. Hwang, "BSN-Care: A secure IoT-based modern healthcare system using body sensor network," *IEEE Sensors J.*, vol. 16, no. 5, pp. 1368–1376, Mar. 2016.
[28] M. Lee and X. Han, "Complex window query support for monitoring streaming data in wireless body area networks," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1710–1718, Nov. 2011.
[29] J. Halamka. (2011). *The Cost of Storing Patient Records*. [Online]. Available: http://geekdoctor.blogspot.tw/2011/04/cost-of-storing-patient-records.html
[30] W. Wang and L. Ying, "Data locality in MapReduce: A network perspective," *Perform. Eval.*, vol. 96, pp. 1–11, Feb. 2016.
[31] F. O. Catak and M. E. Balaban, "CloudSVM: Training an SVM classifier in cloud computing systems," in *Proc. Int. Conf. Pervasive Comput. Netw. World (ICPCA/SWS)*, Berlin, Germany, 2013, pp. 57–68.
[32] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting asthma-related emergency department visits using big data," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 4, pp. 1216–1223, Jul. 2015.
[33] J. Singh, C. Liddy, W. Hogg, and M. Taljaard, "Intracluster correlation coefficients for sample size calculations related to cardiovascular disease prevention and management in primary care practices," *BMC Res. Notes*, vol. 8, no. 1, pp. 1–10, 2015.
[34] H. Alemdar, T. L. M. V. Kasteren, M. E. Niessen, A. Merentitis, and C. Ersoy, "A unified model for human behavior modeling using a hierarchy with a variable number of states," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 3804–3809.
[35] H. Wu, W. Pan, X. Xiong, and S. Xu, "Human activity recognition based on the combined SVM&HMM," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Jul. 2014, pp. 219–224.
[36] D. Lee and K. Roy, "Viterbi-based efficient test data compression," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 31, no. 4, pp. 610–619, Apr. 2012.
[37] K.-C. Wong, T.-M. Chan, C. Peng, Y. Li, and Z. Zhang, "DNA motif elucidation using belief propagation," *J. Nucleic Acids Res.*, vol. 41, no. 16, pp. e153-1–e153-12, Sep. 2013.
[38] R. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw., Pract. Exper.*, vol. 41, no. 1, pp. 23–50, Jan. 2011.
[39] J. Jung and H. Kim, "MR-CloudSim: Designing and implementing MapReduce computing model on CloudSim," in *Proc. Int. Conf. ICT Converg. (ICTC)*, Oct. 2012, pp. 504–509.
[40] M. Lichman. (2013). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

PRASAN KUMAR SAHOO (SM'16) received the M.Sc. degree in mathematics from Utkal University, India, in 1994, the M.Tech. degree in computer science from IIT, Kharagpur, India, in 2000, the Ph.D. degree in mathematics from Utkal University, and the Ph.D. degree in computer science and information engineering from National Central University, Taiwan, in 2002 and 2009, respectively. He was an Associate Professor with the Department of Information Management, Vanung University, Taiwan, and the Software Research Center, National Central University. He is currently an Associate Professor with the Department of Computer Science and Information Engineering and the Director of the International Cooperation Center, Chang Gung University, Taiwan. His current research interests include big data analytic, cloud computing, and cyber-physical systems. He is an Editorial Board Member of the *International Journal of Vehicle Information and Communication Systems*. He has served as a Program Committee Member of several IEEE and ACM conferences. He was the Program Chair of ICCT in 2010.

SUVENDU KUMAR MOHAPATRA received the B.Tech. degree from Biju Pattnaik University, India, in 2008, and the M.Tech. degree from IIIT in 2010. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Division of Computer Science and Information Engineering, Chang Gung University, Taiwan. His research interests include the areas of big data analysis with cloud: medical big data analysis, prediction, optimization, and machine learning.

SHIH-LIN WU (M'15) received the B.S. degree in computer science from Tamkang University, Taiwan, in 1987, and the Ph.D. degree in computer science and information engineering from National Central University, Taiwan, in 2001. Since 2000, he has been with the Department of Computer Science and Information Engineering, Chang Gung University, where he has been a Full Professor since 2015 and the Chairman since 2016. His current research interests include mobile communications, wireless networks, wireless ad hoc networks, and intelligent robots. He is a member of the Phi Tau Phi Society. He serves as a member of the Editorial Board of *Telecommunication Systems*, *Journal of Positioning*, and *ISRN Communications*. He was a Guest Editor of the *International Journal of Pervasive Computing and Communications* in 2007, the Program Chair of Mobile Computing 2005 and the International Workshop on Data Management in Ad Hoc and Pervasive Computing 2009, a Co-Chair of International High Speed Intelligent Communication 2009 and a Co-Chair of International Symposium on Bioengineering 2011, the General Chair of Mobile Computing 2012, a Co-Chair of International High Speed Intelligent Communication and International Conference on Computational Problem-Solving 2013, the Special Session Chair of International Conference on Advanced Robotics and Intelligent Systems 2014 and the Special Session Chair of International Conference on Telecommunication Systems Management 2014, and the Program Chair of International Computer Symposium 2016. Several of his papers have been chosen as Selected/Distinguished papers in international conferences.