International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-4 November 2020

Designing A Cloud-Based Framework using Data Mining for Healthcare Services in Remote Areas

Vijesh Kumar Patel, Jyoti Prakash Singh

Abstract: ***Advancement is recently made inthe medicinal field producesnew innovative technologies to the healthcare sector and medical services. Access to quality healthcare is a major problem in remote areas, with a doctor-to-patient ratio as high as 1:20,000 which is far above the recommended ratio of the World Health Organization (WHO). This has been antagonized by a lack of access to critical infrastructures such as the health care facilities, roads, electricity, and many others. To watch basic medical parameters for identifying the abnormalities within the first stage of chronic diseases need regular interval hospital visits, which can be a comparatively costly and time-consuming process. Rare availabilities of doctors or medical centers, ignorance of the people, and proper care at the right time are the prime causes of great medical concern, which leads to unexpected death. This work is an attempt to solve basic health problems and take advice from registered medical experts for the betterment of the targeted community.Rapid development in the cloud environment, health care services are reasonable to the people in remote areas. It is necessary to predict the disease and connect with the doctor to get an early diagnosis of disease. The imperative goal of the paper is to develop a cloud-based framework using data mining to enhance healthcare in remote areas. The cloud-based framework is designed and simulated by using Matlab R2018b. Fast Search-Growing Self Classifier (FS-GS) data mining classifier is developed to separate the data from the cluster to correlate the symptoms of patients with specialists. The classifier parameters like accuracy, precision, recall, sensitivity, and specificity are analyzed to compare the efficiency of the proposed algorithm with the other data mining algorithms like Naive Bayes, Random Forest, K Nearest Neighbor, and Support Vector Machine Linear. The proposed FS-GS data mining Classifier obtains an accuracy of 92%, the precision of 90.01%, recall of 90.06%, the sensitivity of 94.91%, and specificity of 92.6%. For the effectiveness, the proposed algorithm is compared with the various mining data classification algorithms to showmance of the proposed algorithm. Ultimately, the result shows the proposed algorithm scores higher outputs than all other algorithms in real-time scenarios respectively.***

Keywords: Healthcare Sector, Data Mining, Cloud-Based Framework, Fast Search-Growing Self-Classifier, Matlab.

I. INTRODUCTION

Almost more than half of the total population of rural areas does not have proper access to proper healthcare support due to the shortage of an appropriate number of healthcare facilities and registered doctors.

Revised Manuscript Received on October 10, 2020.
Manuscript Received On October 06, 2020
Vijesh Kumar Patel, Assistant Professor, Department of Computer Science and Engineering, Bakhtiyarpur College of Engineering Patliputra, Patna (Bihar), India. E-mail: vijeshkumarpatel4@gmail.com
Jyoti Prakash Singh, Assistant Professor and Head, Department of Computer Science and Engineering, National Institute of Technology Patna (Bihar), India E-mail: jps@nitp.ac.in

Further comprehensive research indicates that there are certain rural and remote locations in some countries where there are no available hospitals and/or registered doctors to provide required health care facilities.

Thus, against this backdrop, there is an urgent need to understand whether and how to utilize the IT tools to facilitate access to health care support [1]. Lack of access to critical infrastructures such as health care facilities, roads, electricity, and many others makes the situation even worse. Even if these infrastructures are provided, the number of medical practitioners to cater for the growing population of these countries is not sufficient. However, the growing impacts of telemedicine have shown some positive effects in the healthcare delivery system, particularly in developing countries. Henceforth, Information and Communication Technology (ICT) can be used to drive a sustainable and veritable health care delivery system through the introduction and promotion of Virtual Clinics and integration of various health information systems such as Electronic/Mobile Health and Electronic Health Record systems into the healthcare industry in the developing countries [2]. Massive amount of data has to be generated by the healthcare industry traditionally and historically, driven by keeping these records, compliance & restrictive needs, as well as patient caretaking and various other services. These types of data when managed electronically are known as Heath Data Sets. As they are massive and complicated too that they are difficult to manage by already existing code as well as information management tools and ways. The quantity of data is overwhelming not solely due to its volume however additionally due to the range of knowledge varieties and therefore the speed at that it should be managed [3]. Health care monitoring is an effective way for diagnosis, treatment, and prevention of most of the diseases at the early stage. In the modern scenario, advanced technologies like wearable devices, clouds, etc. have made the monitoring or caring of patients simple and flexible. When health parameters data/accuracy is the prior aspect of the system [4]. The advent of mobile health services is becoming a major improvement for patients. The development and Quality of evaluation of a mHealth solution for healthcare professionals in remote areas have to be improved [5]. Recently, the paradigm of online medical prediagnosis has emerged and been recognized as a promising solution to the lack of health professionals in rural areas. Its core idea is combining cloud computing and machine learning techniques for medical automation, such as automated diagnosis and analysis, which will reduce doctors' workload and free them up for more undiagnosed patients.

Therefore, for governments, the online medical prediagnosis scheme is an opportunity to improve the healthcare environment in rural areas [6]. The system should act as a picture centered real-time messaging system to enable the people to order medication online and get it delivered to their residences using cloud computing [7].

Several advantages cloud computing offers to customers in several industries e.g. low-cost services to its customers. However, unlike other industries, the healthcare industry is so far unable to make any significant use of cloud computing [8]. Cloud computing is a structure of resources using different applications. To offer favorable and quick network services, users via the internet server [9] present a new type of cloud computing association, which includes a large number of processors, high-speed networks, memories, and various devices. Various benefits that the cloud offers and presents the challenges that are most prevalent in realizing the full automation of the healthcare system. The integration of machine learning techniques for processing health data in the cloud provides quality healthcare and modernize the healthcare system [10].It is very important to consider the possibility of prospecting useful knowledge form the stored data. The evaluation of the hospital morbidity, prediction through different data mining methods on ambulatory, and hospital procedure records obtained from the databases. The method consists of performing predictive data mining by applying supervised learning algorithms on a regression problem [11]. Cognitive analysis using a data-mining tool is to simulate human thinking into an automated model. The ability to solve problems without the assistance of experts is used in healthcare services. Cognitive analysis can solve complicated problems without the intervention of a medical person [12]. A process is to be identified for reliable health data from online resources and process the data to enable usage [13]. Data mining can be considered as the extraction of raw or useless data from huge databases, many applications like healthcare systems, market analysis get advantages by such mined data, and they came to know how to extract useful data from a big amount of data. This extracted data is most useful to customers [14]. Data mining and Data warehousing is an imperative part of exploring and is realistically worn in diverse domains resembling funding, quantifiable research, teaching, retail, marketing, health care, etc. Many researchers have been systematically been reviewed and surveyed in health care, which is an active interdisciplinary area that is the extent of data mining [15].

II. PROBLEM STATEMENT

Medical health care has been recently increasing concentration and reputation. In machinery resembling atomic, biomedical procedure, therapeutic imaging, and therapeutic records of calm, a huge quantity of health records are produced each day due to advances. Organizations of health care in large volumes of information are generated and collected daily. Data mining is developed day by day in modern existence as new information equipment. It is a process of extracting hidden information from a large, incomplete, noisy data. Data mining is a type of declaration that sustains the method daily. By making inductive interpretation in a highly computerized way, that uses a data warehouse, and then the impending patterns are digging out and making accurate decisions that can analyze the original data to help them to analyze. Its assignation with health care and medical data, data mining has started.

Databases that store health care information, like patient records, which are called Medical data. Plenty of such health records are accumulated in electronic outlines with the development of Information Technology. These databases enclose a large volume of data. Like exponentially the digitized data has amplified, enlarge in the number of records and the catalog mandatory to accumulate. In health care and the remedial ground is persistent and it has several functions like the discovery of scheme in fitness cover, its main purpose is to give appropriate best medical hospital solution at a lower cost like that better medical solution to patients, it another application is to exposure and reason of diseases for patients, and recognition of capable remedial cure methods as a perspective of data mining. However, inthe case ofthe developing countries, the remote areas have no easy access to the health care center or hospitals, it is necessary to provide health care to the people living in the remote areas. Nowadays, due to developments in the cloud environment, health care services are affordable to the people in remote areas. Yet, there is a lag in efficient correlation from symptoms to diseases and diseases to the specialist doctors. It is necessary to predict appropriate diseases based on the symptoms and then display the list of specialist doctors in nearby areas. Currently, all kinds of data related to patients, doctors, symptoms, and diseases are available; it can be used to perform different analyses, to improve the healthcare system by improving the data mining techniques. The research paper is organized in different ways, detailed literature survey in section 3, experimentation, proposed methodology in section 4, result discussion in section 4, and followed by research conclusion in section 6 respectively.

III. LITERATURE SURVEY

As there are a population aging and a decrease in family structures, providing health care to elderly people and the people in remote areas have become a difficult task. The development of mobile internet technology, cloud-computing technology, sensor technology had become a hot topic in mobile health care**.** Thehealth care system using these technologies can be made available to the customers anywhere at any time. The discussion on several techniques in healthcare is discussed below. Inderpreet Singh et al [16] suggested a model of grouping adaptable e-healthcare services administration framework dependent on Cloud Computing. The paper prescribed a model of planning adaptable e-healthcare services administration framework dependent on distributed computing. Unnati Dhanaliya et al [17] presented an E-Health care system by using cloud computing and web services. The use of cloud computing has made remote monitoring and controlling possible. It provided an automatic update of the measured parameter of the patient as well as send alert mail by using SMTP (Simple Mail Transfer Protocol).

Kayo Monteiro et al [18] suggested a health care architecture using IoT for data acquisition, fog for data pre-processing and short-term storage, and cloud for data processing, analyze and long-term storage.

Also, described the main challenges to provide an e-health application with high availability, high performance, and accessibility, at low deployment and maintenance cost. Fekadu workneh et al [19] implemented a cloud-based health care system for storing, retrieving, and updating patient's health records from the Central cloud database server (Dive HQ). An authentication server is also there to filter unauthorized users from accessing the site and to grant access for those authorized users. System development is based on tools like Net Beans IDE, MySQL, and Apache Tomcat. Kuningan Plathong et al [20] presented a conceptual framework on the integration of the Internet of Things with Health Level 7 protocol to support real-time healthcare monitoring by using Cloud computing. The conceptual framework concentrated to help elderly people and ensured people check health care themselves anywhere anytime by using the medical device in the Internet of Things. This real-time storage to Cloud computing is done with JSON language. Therefore, public health and hospitals can use the information for treating patients or advise about healthcare through web service with XML language according to Health Level 7 standards. Said El Kafhali et al [21] studied models and showed to reduce computing resources cost while guaranteeing health requests performance constraints-particularly response time of accessing medical data stored in a fog-cloud environment. To address this issue, the paper proposed a queuing model to predict the minimum required number of computing resources (both fog and cloud nodes) to meet the Service Level Agreement (SLA) for response time. The verification and cross-validation of the analytical model through discrete event simulation was performed. The Obtained results from analytical models showed that the model could correctly and effectively predict the number of computing resources needed for health data services to achieve the required response time under different workload conditions. Maithilee Joshi et al [22] developed a novel, centralized, attribute-based authorization mechanism that uses Attribute-Based Encryption (ABE) and allows for delegated secure access to patient records. This mechanism transfers the service management overhead from the patient to the medical organization and allows easy delegation of cloud-based EHR's access authority to the medical providers. Asif Ahmed Neloy et al [23] proposed a generic architecture, associated terminology, and a classificatory model for observing critical patient's health conditions with machine learning and IBM cloud computing as Platform as a service (PaaS). Machine Learning (ML) based health prediction of the patients is the key concept of this research. IBM Cloud, IBM Watson studio is the platform for this research to store and maintain our data and ml models. For our ml models, we have chosen the following Base Predictors: Naïve Bayes, Logistic Regression, KNN Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and MLP Classifier. For improving the accuracy of the model, the bagging method of ensemble learning has been used. The following algorithms are used for ensemble learning: Bagging Random Forest, Bagging Extra Trees, Bagging KNN, Bagging SVC, and Bagging Ridge.

Ali Nirabi et al [24] presented Mobile cloud computing for emergency health care model (MCCEH) model using a cloud-computing server, MCCEH model providing services related to healthcare in emergency cases and aimed to reduce response time to save the patient's life. When a person is exposed to a health problem or a traffic accident occurs, the MCCEH model will allow users to search for the nearest medical center or nearest specialists related to a specific specialization and the results will show the availability timetable for every specialist and whether he is available at this time or not, the user will be able to choose specialist / medical center based on previous experiences may able to read previous feedback and opinions. Hosam F. El-Sofany et al [25] proposed a fuzzy model that will be used for developing a cloud-based health application for medical diagnostic.

IV. PROPOSED RESEARCH METHODOLOGY

Rural and remote residents often encounter barriers to healthcare that limit their ability to obtain the care they need. For rural residents to have sufficient access, necessary and appropriate healthcare services must be available and obtainable on time. Even when an adequate supply of healthcare services exists in the community, there are other factors to consider in terms of healthcare access. The proposed method paves a way for the remote residents to access health care services through cloud computing via advanced data mining techniques. The proposed framework combines all the main components of the healthcare system together that are patients in remote areas, doctors, symptoms, and diseases. It provides patients or users to get the required health services through a communication gadget. The patient shall log in to the mobile application and enters the symptoms and location. Based on the symptoms entered by the user, the nearest specialized doctors and hospitals concerning the user's location can be obtained by the user. The system relates entered symptoms to diseases by analyzing historical data maintained by the system. The system analyzes by finding the closest match of the disease corresponding to the symptoms. Then it maps this matched disease to specialized doctors by acquiring details from the database. This mapping is displayed as a set of specialized doctors and their details at the user's end. The patient and book an appointment can choose a specialized doctor or a hospital. On the other hand, doctors can either accept or reject the appointment and an acknowledgment. This is displayed in the user's account accordingly. The patient can visit the doctor and are treated with whom the appointment has been fixed. The doctor enters the actual disease being suffered by the patient into the system through his account as per the treatment. The entry made by the doctor updates the overall cloud database of the system that further may fortify the availability of data and enrich it. With more entries, the analysis of this database also is enhanced. The patient can access his data while the doctor manages and maintains the patient records that he can use for his purpose. The system updates its historical database daily.

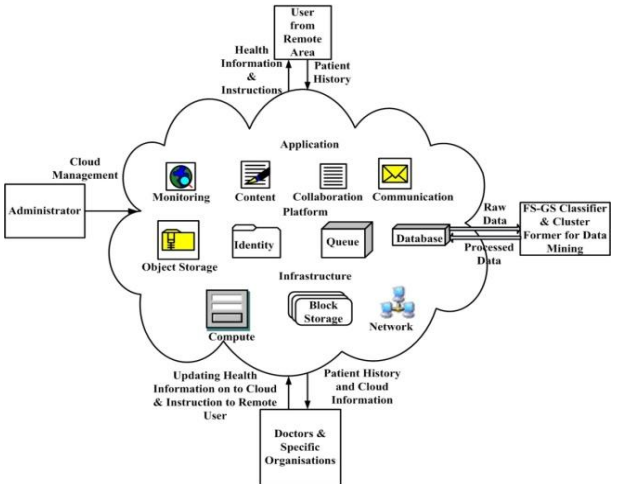

Fig. 1 Architecture of Proposed Methodology

The proposed methodology using a cloud interface that puts on view the results of the data mining techniques from the back end has shown in Figure 1. It elaborates on the integration of data mining techniques into the cloud-based healthcare application and gathers useful results. The proposed FS-GS data-mining tool concentrates on clustering with classification for analysis purposes by reduced computation time using fast search in clustering and the advanced variant of self-organizing maps for classification with improved interpretability. When a particular scenario is clicked, it fetches data from the related database table by converting it into the specified format. By using this data file, we use the proposed data mining techniques, i.e., clustering and classification to display results. These results enable different organizations and governments to improve the healthcare system. The Healthcare System may use any deployment model based on administrator requirements.

Cloud Computing: Cloud Computing delivers computing services such as servers, storage, databases, networking, software, analytics, intelligence, hardware, virtual desktop, applications and software platform, and more, over the Cloud (Internet) to the health care service framework. Each has its unique importance in cloud architecture. The cloud environment provides an easily accessible online portal that makes handy for the user to manage the resources.The types of cloud services include Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The selection of the service is administrator oriented.

Fast Search - Growing Self Classifier Data Mining Technique: Data mining finds valuable information hidden in large volumes of data. It is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. Figure 2 shows the FS-GS data-mining tool in detail. The techniques used are responsible for finding the patterns by identifying the underlying rules and features in the data. The FS-GS data mining technique gives attention to the clustering and classification process. The FG-GS Data Mining tool is a fast search & GSOM based interpretable classifier and cluster former. FS-GS interpretable classifier is based on GSOM (Growing Self Organizing Maps) algorithm. The GSOM is further based on SOM used in the earlier data mining techniques, which are used to identify the nearest prototypes for generating fuzzy rules. The proposed algorithm, which is Growing Self Organizing Maps based interpretable classifier, depends on the cognitive process of human thinking for classification task by forming prototypes, build discrete rules under each concept to be able to identify specific objects, and performs classification on the decided clusters formed by a fast search-clustering algorithm. The fast search-clustering algorithm is based on the natural neighbor-clustering algorithm, which emerged to automatically widen the searching range, to ensure all the data points are in a naturally stable state, for data mining. The fast search of a natural neighbor algorithm is applied to the classifier results for fast search of natural neighbors by quickly finding the natural characteristic value and natural neighbors. It depends on the most distant neighbor in the data set. We can shorten the time of finding the natural neighbors of each data point with this algorithm.

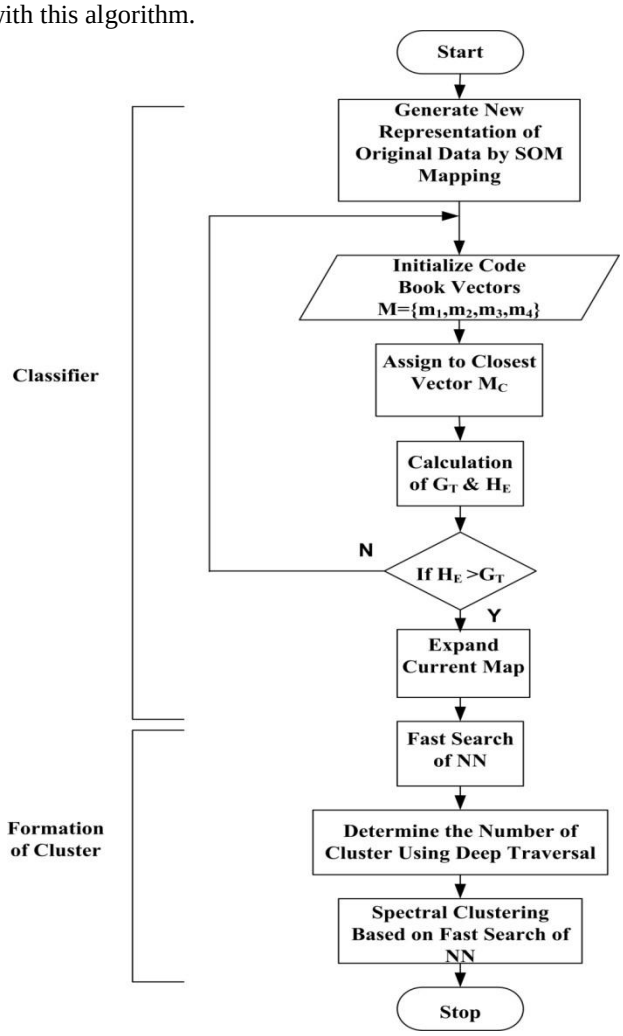

Fig. 2 Flow Chart for FS-GS Data Mining Tool

Clustering: Depending on the nearest neighbor list of the most remote data points, the rankings of all the farthest points of data in the list should be obtained. Then regard the largest ranking as the natural characteristic value. $\sup_k$, thefast search algorithm finds the natural characteristic value. For a data set X, x_i is the farthest neighbor in X .A Euclidean metric is used for distance calculation between x_i and x_j. Euclidean distance is given by,

$$d_{x_i,x_j} = \sqrt{\sum_{m=1}^{M}(x_{im} - x_{jm})^2} \quad (1)$$

Where m is represents dimension and M refers to the number of elements in the data set. It is necessary to design a function to seek the nearestneighbor x_j x_i. The find-ranking method is needed to be designed to find the ranking R of x_i the nearest neighbor list x_j. Finally, return the maximum value R and set R as the characteristic value $\sup_k$. The end of the algorithm is the last row in the neighbor list of data sets, which consists of the most remote data point x_i, x_j is the nearest neighbor of x_i. R is the ranking of x_i, and the maximum value R is set as the natural characteristic value of X. The nearest neighbor (NN), a search function is required to get $\sup_k$,the nearest neighbors of each data point in X.

Determining the Number of Clusters based on Deep Traversal: For finding the proper number of clusters, a method must be put forward by performing a deep traversal of all data points in the same cluster. Firstly, it selects one data point in the set of data and detects whether it has been accessed already. If it is not accessed before, the data point will be updated as an accessed one. On the contrary, a selection process goes ahead to verify whether there are unvisited points in the set of data. Repeating this process until the whole point is visited is needed. Finally, this process will get to an end and the number of clusters is estimated. In this algorithm, all points of data in the same extended area are marked with the variable C. However, the different input orders of data points have a huge impact on the performance of clustering. To solve this problem, the concept of Natural Full Neighborhood is provoked. The data points are expanded in the Natural Full Neighborhood to reach an ideal result in clustering. The Natural Full Neighborhood is described as follows. For each point x_i in the set of data X, the full neighbors of the point x_i comprise of its natural neighbors and natural inverse neighbors represented as SNN. It is defined as follows,

$$SNN(x_i) = \{NN(x_i) \bigcap RNN(x_i)\} \quad (2)$$

Where $NN(x_i)$ denotes the natural neighbor list of x_i, $RNN(x_i)$ represents the natural inverse neighbor list of x_i .For each point x_i in the data set X, the natural full neighborhood x_i consists of the data point itself and its natural full neighbors. The natural full neighbors are provided as given below,

$$neighbor(x_i, SNN) = \{x_i\} \bigcup \{SNN(x_i)\} \quad (3)$$

Here, the value C represents the number of clusters and is superimposed based on the access status of the data points. It is increased by 1 every time, which means that the natural full neighbors of one point are fully accessed and a cluster is created. When all the points in the data set are visited, superimposing stops here. The final value C is the number of clusters.

Spectral Clustering based on Fast Search of Natural Neighbors: In the traditional natural neighbor algorithm, the entire computation procedure of natural neighbor can be automatically satisfied without any parameters. So the natural neighbor of each data point can be precisely found, the relationship between data points is precisely reflected, and the data structure is exactly determined. However, its search speed is relatively low, which is the greatest challenge. Therefore, focusing on a quickly and accurately determining the natural characteristic value of the data set and betterment the search efficiency, a fast search of natural neighbors is defined.

The clustering part of the FS-GS algorithm contains three main components: Fast search of the natural neighbors, determining the number of clusters on deep traversal, and performing with a spectral clustering algorithm. There is no need to set any parameters and search cost is effectively decreased. By determining the natural neighbor of each point, the similarity function can be redefined. This computes the similarity between natural neighbors so it greatly reduces the cost of computation. Finally, clustering is performed with a spectral clustering algorithm. The similarity between natural neighbors is calculated as shown below.

$$d(x_i, NN(x_i)) = \sum_{t=1}^{m}\sqrt{(x_i - NN_{x_i t_i})^2} \quad (4)$$

$$w(x_i, NN(x_i)) = \exp(\frac{-d^2(x_i, NNx_i)}{\sigma^2}) \quad (5)$$

Where NNx_i are the natural neighbors of x_i, σ is set to a constant value depending on experience.

Classifier: The first step of the classifier is, generating a new representation of the original data, which serves as a basis for forming appropriate groups. GSOM algorithm, which is a variant of SOM (Self-Organizing Maps), is employed here. Growing Self Organizing Maps is chosen because of its ability to form a dynamic and self-organizing representation of data inputs. The new representation has the characteristic that similar objects tend to be mapped to nearby points in 2D space. Specifically, forgiven a set of data D ; at first, four codebook vectors are initialized as $M = \{m_1, m_2, m_3, m_4\}$, where each codebook vector has the same number of properties or attributes with the dataset D. After that, a comparison of an instance x in D with every codebook vector using Euclidean distance is performed and assigned to its closest vector m_c where, $c \in \{1,2,3,4\}$.

After the assignment of a new data instance, x to the existing codebook vector m_c is also represented as the winner node, for reducing the quantization error, the values of the winner node and its neighbors will be updated. Consequently, the winner node and its neighbors will be closer to the assigned data point after the updating process, where t is the time stamp of each step, α is the learning rate set to reduce over each step, and N is the set containing neighbors of m_c.

$$mi(t+1) = \begin{cases} m_i(t); i \notin N_{t+1} \\ m_i(t) + \alpha(t) * (x - m_i(t)); i \in N_{t+1} \end{cases} \quad (6)$$

Practically, there exists the case that most of the data instances are assigned to a single node in the map, which is known as the under-representation of the data. In such a case, new nodes will be grown from the existing node in other to relax the problem and represent the data more probably. Specifically, the maximum error H_E is set to keep track of the highest error of each node on the map M. Where the error, E_C of a node, m_c is the accumulated distance between it and the assigned data point x.

$$E_C(t+1) = E_C(t) + \|x - m(t)\| \quad (7)$$

$$H_E = \arg\max_i (Ei) \quad (8)$$

After fitting all training data D to get their new depiction by using a growing self-organizing maps algorithm, the result of the above step is a smooth self-organized map M containing i codebook vectors $M = \{m_1, m_2, m_3, ...m_g\}\ (g \geq 4)$. Each codebook vector is the representation of a Voronoi region in the original data space. The generated self-organizing map M has the characteristic that codebook vectors that denote groups of similar objects that lie adjacently to each other. To automatically find out those semantic groups of codebook vectors, a simple but effective fast search to cluster the map M is used. Let $C = \{C_1, C_2, C_3, ..., C_k\}$ be the set of k clusters in M, for any two different clusters C_j and $C_{j'}$, we have,

$$C_j \cap C_{j'} = \phi \text{ if } j = j' \text{ and } M = \bigcup_{j=1}^{k} C_j \quad (9)$$

Furthermore, for each cluster C_j, the center C_j is defined as,

$$Vj = [V_{j1}, V_{j2}, ...V_{jl}] \quad (10)$$

Then we have the general formulation for comparison of the dissimilarity between a codebook vector, $m_i \in M$ and a cluster center V_j described below as the Euclidean distance between m_i and V_j.

$$dis(m_i, v_j) = \|m_i - v_j\| \quad (11)$$

Based on equations (10) and (11) the fast search algorithm-clustering algorithm aims to minimize the following objective function,

$$J(U, D) = \sum_{i-1}^{g} \sum_{j=1}^{k} U_{i,j} \times dis(m_i, v_j) \quad (12)$$

Where $U = [U_{i,j}]_{g \times k}$ is the partition matrix.

The ending part is to perform the classification task; building decision rules for each cluster $C_j \in C$ using the CART (Classification and Regression Tree) algorithm, which is a well-known binary decision tree-learning algorithm, is essential here.This strategy tends to reduce the computation cost and increasing the intelligibility of the model by providing a very less number of shorter decision trees while it still preserves a high performance. Specifically, provided a cluster $C_j \in C$ that composes a set of original data instances that are similar to each other, starting from the entire set of data instances belonging to C_j which is the root node the data will be split on the feature that results in the largest information gain (IG). To split the nodes at the most informative features, the objective function is to maximize the information gain at each split as defined by the following formula,

$$I_G(D_{p,att}) = I(D_p) - \left[\frac{N_{left}}{N_p} I(D_{left}) + \frac{N_{right}}{N_p} I(D_{right}) \right] \quad (13)$$

Where attis the feature to perform the split, D_p and D_{left}, D_{right} are the dataset of the parent, left and right child nodes respectively, I is the impurity measure, N_p is the total number of samples at the parent node, and N_{left}; N_{right} are the number of samples in the left and right child node. The difference between the impurity of the parent node and the sum of the child node impurities the lower the impurity of the child nodes is simply the information gain.

The Gini index (GI) is used as the measure of impurity. Intuitively, the Gini index can be defined as a criterion to reduce the probability of misclassification and given as the below formula where $P(i|t)$ is the proportion of the samples which belong to class C for a particular node t.

$$I_{GI}(t) = \sum_{i=1}^{c} P(i|t)(-P(i|t)) = 1 - \sum_{i=1}^{c} P(i|t)^2 \quad (14)$$

In an iterative process, this splitting procedure at each child node can be repeated until the leaves are pure. This shows that the samples at each node all belong to the same class. The system operates with binary classification, therefore $C = 2$. A detailed explanation of the FS-GS classifier is provided above.

V. EXPERIMENTAL ANALYSIS AND RESULT DISCUSSION

The proposed method can beimplemented in the working platform of Matlab R2018b with the following system specification. The Matlab R2018b software is used for the design of the proposed algorithm. The following system configuration is utilized to execute the proposed paradigm in the research work.

Table I: Simulation System Configuration

Software Platform	MATLABR2018b
Operating System	Windows 10 Home
Processor	Intel Core i5 @ 2.4GHz
Memory	8 GB

Table 1 shows the system configuration on simulation, as stated above. The MATLABR2018b is used as the software platform for simulation. The OS used is Windows 10 Home with Intel core i5 operating at 2.4 GHz and 8 GB memory storage. There may be limited variation towards other system configurations. The design of the cloud framework using datamining is exhibited. To achieve the successful prediction for symptoms and doctors, the proposed FS-GS data mining classifier is used. The proposed data-mining algorithm chooses the exact data from the cluster. The classifier parameters like accuracy, precision, recall, sensitivity, and specificity are analyzed. The number of predictors used for analysis is 50. The obtained outputs are shown below and the graphical representation of each parameter has been deliberated below. **Accuracy:**Accuracy is the rationumber of correctly classified instances to the total number of instances. The formula for calculating accuracy is shown in Equation 9 where, TP = True Positive, FP = False Positive, FN= False Negative, TN= True Negative.

$$\text{Accuracy} = \text{TN} + \text{TP} \div \text{TP} + \text{TN} + \text{FN} + \text{FP} \quad (15)$$

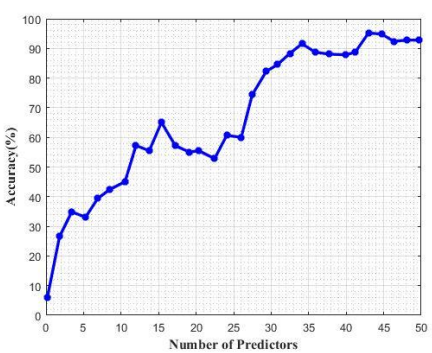

Fig. 3 Graph for Accuracy

Figure 3 shows the graph result, which illustrates the number of predictors about the patient's symptoms and the specialist list against the accuracy rate.The number of predictions and the correct instances of predictions is calculated to plot the graph. The graphical representation shows that there is a gradual increase from range 0 to 43 predictors and it steadily decreases at the range 50. The accuracy thus obtained for the number of predictions is 92%. It shows the predictions like symptoms of patients, doctors list are eventually predicted correctly.

Precision: Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances. Precision is the ratio of actual true predicted instance out of total true instance.

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP} \quad (16)$$

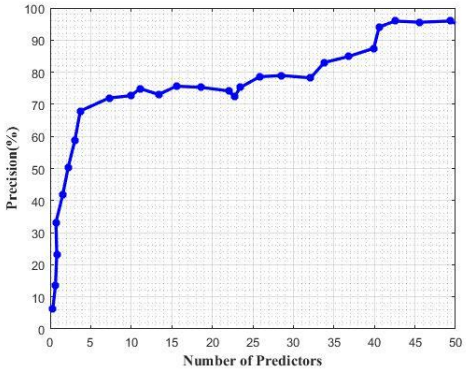

Fig. 4 Graph for Precision

Figure 4 shows the graph result, which illustrates the number of predictors about the patient's symptoms and the specialist list against the precision level. From the retrieved predictors, the number of positive prediction data is taken to plot the graph. At the fifth predictors, there is a gradual increase where the precision level increases up to 70% after that the number of positive predictions goes on increasing. Therefore, the graph clearly shows that the designed framework can provide a positive result of precision up to 90.01%. The symptoms against the patients are correctly correlated to get positive predictors.

Recall: The recall is a ratio of actual true instance out of all true items. The recall is the fraction of the relevant documents that are successfully retrieved.

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \quad (17)$$

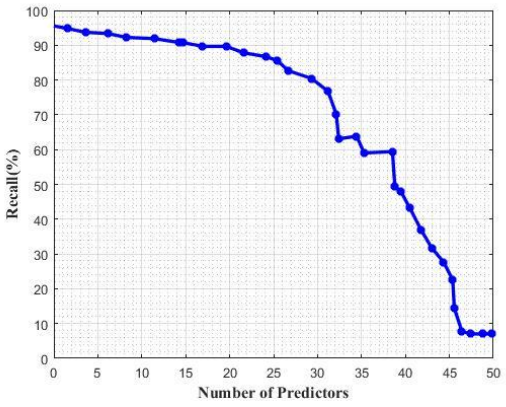

Fig. 5 Graph for Recall

Figure 5 shows the graph result, which illustrates the number of predictors about the patient's symptoms and the specialist list against the recall.

The recall represents the maximum level of true predictions from the positive predictor. The above graph represents how the recall has been increased at the range of zero and decreased at the range of 50. This shows the newly designed framework helps in health care by providing true instances for the patients. The recall rate for the number of predictors is 90.06%.

Sensitivity: Sensitivity is the proportion of actual positives, which are correctly identified as positives by the classifiers.

$$\text{Sensitivity} = P/P + N \quad (18)$$

Where P represents positive predictions and N represents the negative predictions.

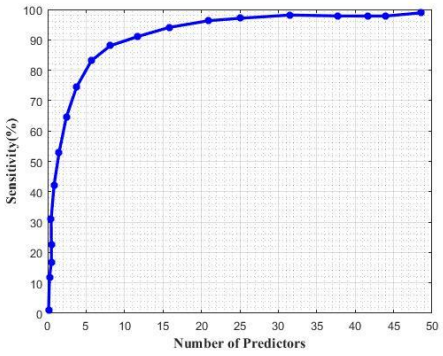

Fig. 6 Graph for Sensitivity

The graph result, which illustrates the number of predictors about the patient's symptoms and the specialist list against the sensitivity, has been shown in figure 6. The graphical representation illustrates that there is a steep increase from the range 0 to 5 it attains a rate of 80%, then it steadily increases up to the range of 50. The sensitivity indicates the number of true instances from the prediction. The sensitivity rate for the number the predictors is 94.91%. The calculated result indicates how well the designed framework works.

Specificity: Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified.

$$\text{Specificity} = N/N + P \quad (19)$$

Where P represents positive predictions and N represents the negative predictions.

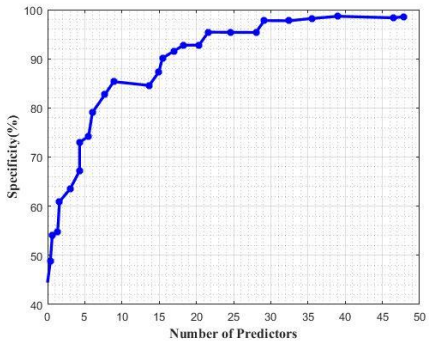

Fig. 7 Graph for Specificity

Figure 7 shows the graph result, which illustrates the number of predictors about the patient's symptoms and the specialist list against the specificity. The graph represents that there is a steady increase from the range of 0 to 12 and then there is a decrease within the range of 15, then it goes on increasing up to the 46 predictors. Specificity is the number of negative instances provided by the number of predictions. The negative instances show how well the predictions go on successful. It represents the symptoms are correctly correlated with the specialist. The specificity rate for the number of predictors is 92.6%.

Table II: Classifier Parameters

Parameters	Percentage
Accuracy	92%
Precision	90.01%
Recall	90.06%
Sensitivity	94.91%
Specificity	92.6%

The optimized values of the classifier parameters using the proposed are represented in Table 2. The parameters like accuracy, precision, recall, sensitivity, and specificity are calculated. The accuracy of the proposed algorithm is 92%, precision is 90.01%, recall is 90.06%, sensitivity is 94.91%, and specificity is 92.6%.

Comparison Stratagems: To evaluate the overall performance of the proposed algorithm, it is compared with other algorithms like Naive Bayes, Random Forest, K-Nearest Neighbor(KNN), Support Vector Machine (SVM) Linear algorithms. The performance results are been discussed below:

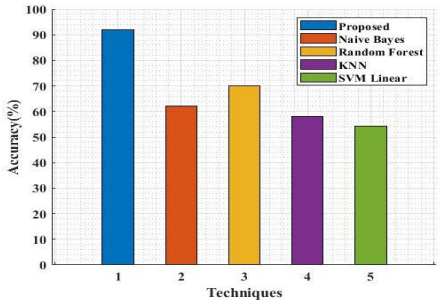

Fig. 8Comparison Graph for Accuracy

The comparison of accuracy on algorithms like the proposed algorithm, Naive Bayes, Random Forest, K-Nearest Neighbor(KNN), Support Vector Machine (SVM) Linear algorithms is shown in figure 8. The accuracy rate increases by 92% for the proposed algorithm whereas, the accuracy rate for other algorithms is as follows, Naïve Bayes is up to 62%, Random Forest is up to 69.9%, KNN is up to 58.1% and SVM Linear is up to 54.2% respectively. The Accuracy comparison result shows that the proposed algorithm is better than other algorithms.

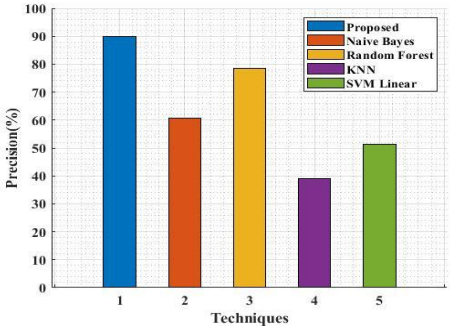

Fig. 9 Comparison Graph for Precision

Figure 9 shows the comparison of precision with the proposed and other algorithms. The result shows that the precision rate for proposed algorithms is 90.01%, for Naïve Bayes precision is 60.69%, for Random Forest it is 78.37%, for KNN it is 39.03%, and for SVM Linear precision rate is 51.24% respectively.

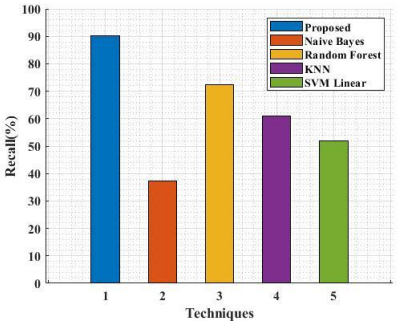

Fig. 10Comparison Graph for Recall

Figure 10 shows the comparison result for recall. The recall rate for the proposed algorithm is 90.06% and for other algorithms are Naïve Bayes recall is 37.4%, for Random Forest is 72.5%, for KNN is 60.87%, and for SVM Linear is 51.98% respectively.

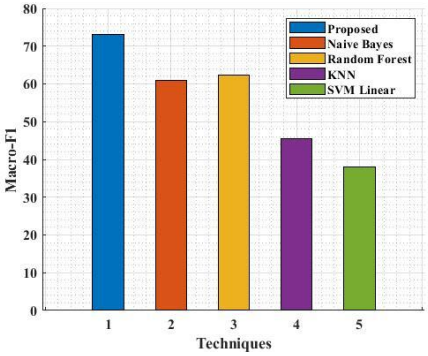

Fig. 11 Comparison Graph for Macro F1

The F1 score is defined as the weighted harmonic mean of the test's precision and recall. It is calculated by,

$$F1\ measure = 2(Precision * Recall)/(Precision + Recall) \quad (20)$$

The Macro F1 measure represents the measure of test accuracy, by considering both precision and recall. A comparison of Macro F1 measure with proposed and other algorithms is shown in figure 11. The proposed algorithm shows a Macro F1 rate of 73% whereas, the Macro F1 rate for other algorithms is as follows, for Naive Bayes up to 61%, for Random Forest up to 62.3%, for KNN up to 45.6%, and SVM linear up to 38.1% respectively. The Macro F1 comparison result shows that the proposed algorithm is better than other algorithms.

Table III: Comparison Results between Algorithms

Techniques	Accuracy %	Precision%	Recall%	Macro-F1%
Proposed	92	90.01	90.06	73
Naïve Bayes	62	60.69	37.4	61
Random Forest	69.90	78.37	72.5	62.30
KNN	58.10	39.03	60.87	45.60
SVM Linear	54.20	51.24	51.98	38.10

The comparative experiments were repeated to compare the classifier's parameters with the other algorithms like the Naïve Bayes algorithm, Random Forest algorithm, K Nearest Neighbour (KNN) algorithm, Support Vector Machine (SVM) Linear algorithms. The results are shown in Table 2. The result indicates that the proposed algorithm is the best algorithm used to design the cloud-based framework for health care. It is because the algorithm has the best accuracy than other algorithms (refer table 2).

VI. CONCLUSION

With the advent of new technologies in the digital health field and the growth of the human population, medical surveillance systems have become of paramount importance. In this research, a cloud-based framework is designed from where the data can be collected. The data like the symptoms of the patients is correlated with the availability of doctors in that field. The list of the specialist doctor will be displayed by using the data mining technique where the data are gathered by using the FS-GS classifier and cluster. By using this method, one can easily access the specialist according to their requirements. The classifier parameters like accuracy, precision, recall, sensitivity, and specificity are calculated for 50 predictors. The efficiency of the obtained parameters rate compared with the other data mining algorithms like the Naive Bayes algorithm, Random Forest Algorithm, K-nearest neighbor algorithm, Support Vector Machine Linear algorithm is been given below:

- ❖ Accuracy results show that the proposed algorithm is 92%, whereas for data mining algorithms like the Naive Bayes algorithm, Random Forest algorithm, K-nearest neighbor algorithm, Support Vector Machine Linear algorithm accuracy ranges in between 50 to 70%.
- ❖ Precision results show that the proposed algorithm is 90.01%, whereas for data mining algorithms like the Naive Bayes algorithm, the Random Forest algorithm, K-nearest neighbor algorithm, Support Vector Machine Linear algorithm precision ranges in between 45 to 80%.
- ❖ Recall results show that the proposed algorithm is 90.06%, whereas, for data mining algorithms like the Naive Bayes algorithm, the Random Forest algorithm, K-nearest neighbor algorithm, Support Vector Machine Linear algorithm recall ranges in between 35% to 75%.
- ❖ Macro-F1 results show that the proposed algorithm is 73%, whereas for data mining algorithms like the Naive Bayes algorithm, the Random Forest algorithm, K-nearest neighbor algorithm, Support Vector Machine Linear algorithm macro f1 ranges in between 35 to 65%.

The above results illustrate that the proposed FS-GS data mining classifier algorithm produces better performance when compared with other data mining algorithms like the Naive Bayes algorithm, the Random Forest algorithm, K-nearest neighbor algorithm, Support Vector Machine Linear algorithm. In the future, the work can be extended by enhancing the cloud framework with artificial intelligence by connecting it with smartwatches. Therefore, the smartwatches will help the patients by providing better datasets like symptoms, specialists, appointments, prescriptions, and improve the accuracy of prediction. Hence, the overall research outcome proves that the healthcare sector can monitor the disease symptom and patients to correlate effectively.

By utilizing this prediction method, which takes less time to correlate the symptoms and desired specialist, therefore the end-user can easily progress appointments in real-time conditions respectively.

REFERENCE

1. L.Muhammad Hassan Bin Afzal, J.D.Daniel Cravens,"Technology-based public policy interventions to enhance access to health care support: A study in the context of rural Bangladesh",*IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* pp.21 - 23 Dec 2017.
2. Nasir Faruk, Nazmat T. Surajudeen-Bakinde, Abdulkarim A. Oloyede, "On green virtual clinics: A framework for extending health care services to rural communities in Sub-Saharan Africa",*2017 International Rural and Elderly Health Informatics Conference* 978-1-5386-6016-4.
3. Shalu Gupta, Dr. Pooja Tripathi, "Big data lakes can support better population health for rural India-Swastha Bharat*", 2016 1st International Conference on Innovation and Challenges in Cyber Security (ICICCS2016)*, 978-1-5090-2084-3/16/2016 IEEE.
4. Mohammad Jabirullah, Rakesh Ranjan, Mirza Nemeth Ali Baig, Anish Kumar Vishwakarma, " Development of e-Health Monitoring System for Remote Rural Community of India",*2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)* 978-1-7281-5475-6/20/IEEE.
5. Bruno, Silva, Joel J. P. C. Rodrigues, Andre Ramos," A Mobile Health System to Empower Healthcare Services in Remote Regions", *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)* 978-1-7281-0402-7/19/IEEE.
6. Wei Guo; Jun Shao; RongxingLu; Yining Liu; Ali A. Ghorbani," A Privacy-Preserving Online Medical Prediagnosis Scheme for Cloud Environment", vol.4, 2016, pp. 48946 – 48957.
7. Sandesh Chinchole; Aditya Kulkarni; Laksh Matai; Chintan Kotadiya, " A real-time cloud-based messaging system for delivering medication to the rural areas",*IEEE International Conference on Intelligent Sustainable Systems (ICISS)*, Dec. 2017, pp.7-8.
8. Mayank Singh, P. K. Gupta, Viranjay M. Srivastava," Key challenges in implementing cloud computing in the Indian healthcare industry",*2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech).*
9. Adesh Kumari, Vinod Kumar, M. Yahya Abbasi, Saru Kumari, Pradeep Chaudhary, Chien-Ming Chen, "CSEF: Cloud-Based Secure and Efficient Framework for Smart Medical System Using ECC",*IEEE Access* vol. 8,09 June 2020, pp. 107838 – 107852.
10. S Vidya Priya Darcini., Deva Priya Isravel, Salaja Silas," A Comprehensive Review on the Emerging IoT-Cloud based Technologies for Smart Healthcare", *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 23 April 2020.
11. Leonardo Silva Vianna, Raul SidneiWazlawick," Data Mining for Hospital Morbidity Forecasting",*2020 IEEE International Conference on Software Architecture Companion (ICSA-C)*;March 2020,pp.16-20, 978-1-7281-4659-1/20
12. Suneeta S. Raykar, N.Vinayak," Cognitive Analysis of Data Mining Tools Application in Health Care Services",*2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 27 April 2020.
13. Hong Qing Yu," Extracting Reliable Health Condition and Symptom Information to Support Machine Learning", 09 April 2020, IEEE.
14. Sopan, Ganpat, Sutar," Intelligent data mining technique of social media for improving health care*", 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 11 January 2018, IEEE.
15. Md. Robel Mia, Syed Akhter Hossain, Amit Chakraborty Chhoton, Narayan Ranjan Chakraborty," *A Comprehensive Study of Data Mining Techniques in Health-care, Medical, and Bioinformatics*", 2018 International Conference on Computer, Communication, Chemical, Material, and Electronic Engineering (IC4ME2),20 September 2018.
16. Inderpreet Singh, Deepak Kumar, Sunil Kumar Khatri," Improving The Efficiency of E-Healthcare System Based on Cloud", *IEEE 2019 Amity International Conference on Artificial Intelligence (AICAI)* 978-1-5386-9346-9/19.
17. Unnati Dhanaliya, Anupam Devani," Implementation of E-health care system using web services and cloud computing",*International Conference on Communication and Signal Processing*, April 6-8, 2016, India.
18. Kayo Monteiro, Elisson Rocha, Emerson Silva, Guto Leoni Santos, Williams Santos, Patricia Takako Endo," Developing an e-Health System Based on IoT, Fog and Cloud Computing*", 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)Date of Conference:* 17-20 Dec. 2018, 978-1-7281-0359-4/18.
19. Fekadu worked, Ahmed Adem, Roshni Pradhan, "Understanding Cloud-Based Health Care Service with Its Benefits*",2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE* September 2018.
20. KuntinanPlathong, BoonprasertSurakratanasakul," A study of integration Internet of Things with health level 7 protocol for real-time healthcare monitoring by using cloud computing", *The 2017 Biomedical Engineering International Conference (BMEiCON-2017)*, 978-1-5386-0882-1/17.
21. Said El Kafhali, Khaled Salah, Said Ben Alla," Performance Evaluation of IoT-Fog-Cloud Deployment for Healthcare Services",*IEEE 2017 10th Biomedical Engineering International Conference (BMEiCON)*, vol.31 Aug.-2 Sept. 2017.
22. Maithilee Joshi, Karuna P. Joshi, Tim Finin," Performance Evaluation of IoT-Fog-Cloud Deployment for Healthcare Services*", IEEE, 2018 4th International Conference on Cloud Computing Technologies and Applications (Cloud tech)*,pp. 2159-6190/18.
23. Asif Ahmed Neloy, Sazid Alam, Rafia Alif Bindu, Nusrat Jahan Moni, "Machine Learning-based Health Prediction System using IBM Cloud as PaaS",*2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*,Nov. 2018,pp. 26-28.
24. Ali Nirabi, Shihab A. Hameed,"Mobile Cloud Computing For Emergency Healthcare Model: Framework",*2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, 978-1-5386-6992-1.
25. F.Hosam, El-Sofany, Islam A.T.F. Taj-Eddin, " A Cloud-based Model for Medical Diagnosis using Fuzzy Logic Concepts",*IEEE 2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*; 978-1-5386-5261-9/19.