JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

THEIL SEN REGRESSION AND CANOPY HOPKINS STATISTIC CLUSTERING FOR E-HEALTHCARE MONITORING WITH IOT AND BIG DATA

V.Deepa

Research Scholar, Tiruppur Kumaran College for Women,
P.G and Research Department of Computer Science, Tirupur.

Dr K.Rajeswari,

Associate Professor, P.G and Research Department of Computer Science,

Tiruppur Kumaran College for Women, Tirupur

Abstract

New possibilities for E-health care monitoring are provoked by the expansion of Internet of Things (IoT) and Big data, in addition to the pervasive nature of small wearable sensors. The IoT and big data is a paramount issue in numerous domain areas including e-healthcare systems due to its importance. Big data are notably utilized in e-healthcare to ascertain the normal and abnormal patient condition. Numerous research works were introduced for e-healthcare monitoring by many researchers varying from diagnosis to disease recognition and prevention on effective e-healthcare monitoring. Numerous issues like, accuracy, time and error have yet to be conveyed to generate a ductile system for health care monitoring. To address these issues, in this work, a method called Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) for IoT-based healthcare monitoring is proposed. The TSLR-CHSC method is split into three sections, namely, data collection, feature selection and clustering. First, big data comprising of cardiovascular disease dataset acquired from sensors are collected. Next, relevant features with maximum accuracy and minimum time are obtained using the Theil–Sen Estimated Linear Regression Feature Selection model. Followed by which with the relevant features, clustering is performed by means of Canopy Hopkins Statistic Clustering for healthcare monitoring. Here, with the aid of Canopy Clustering determining the cluster tendency to what degree clusters exist in data to be clustered. By this way, an efficient diseased patient health monitoring is carried out with minimal time consumption. For experimentation, a systematic cardiovascular healthcare data is produced utilizing kaggle dataset and medicinal gadgets to foresee the diverse patient levels of disease severity. A detailed comparative analysis is carried out and the simulation outcome ensured the goodness of the TSLR-CHSC method over the compared methods under various aspects.

Keywords: Internet of Things, Big Data, Theil–Sen, Linear Regression, Feature Selection, Canopy Hopkins, Statistic Clustering, healthcare monitoring

1. Introduction

In the recent years healthcare monitoring system in hospitals and health centers has experienced large development and therefore portable healthcare monitoring systems is found to be in rising trends globally. Moreover, the initiation of Internet of Things (IoT) helps the progress of healthcare from face-to-face consulting to telemedicine. Internet of Things (IoT) is an environment where every connected node communicates with other nodes order to transfer essential data for accurate decision making. Smart healthcare system in IoT environment monitor patient basic health signs in real-time.

An IoT-based student healthcare monitoring model was introduced in [1] to check the student vital symptoms and identify both biological as well as behavioral changes via smart healthcare technologies. Here, the vital data were acquired from IoT devices. Moreover, data analysis was performed by means of machine learning techniques for identifying probable risks concerning student physiological and behavioral variations. However, the extent of accuracy was not reduced using IoT-based student healthcare monitoring model.

A PATH2iot framework was introduced in [2] to decompose complex IoT application into micro-operation. Depending on the deployment, PATH2iot distributed set of micro-operation across IoT infrastructure platform while considering the runtime data and control flow dependencies. PATH2iot introduced heuristic model for taking optimal deployment decisions depending on multiple non-functional requirements and selection criteria. But, the energy consumption for data communication was not minimized by PATH2iot framework.

An Energy Efficient Particle Swarm Optimization (PSO) based Clustering (EEPSOC) method was introduced in [3] for efficient cluster head (CH) selection among different IoT devices. IoT devices were grouped into cluster for sensing healthcare data. Followed by which CH was elected by EEPSOC. The elected CH transmitted data to cloud server. Here, the CH was utilized in sending IoT devices data to cloud server via fog devices. Also, an artificial neural network (ANN) based classification model was introduced to diagnose healthcare data in cloud server to recognize disease severity. Though the energy consumption was reduced, time complexity was not focused by EEPSOC method.

An efficient sensor-based data analytics was introduced in [4] for real-time patient monitoring to help hospital and medical staff. The designed mechanism comprised three phases, namely emergency detection, adapting sensing frequency and real-time patient situation prediction. However, sensor nodes failed to avoid repeated collision. The designed mechanism failed to adjust sensing frequency based on available energy beside redundancies between readings at different periods.

1.1 Motivation

For IoTs related to health sciences with big data, healthcare monitoring is a crucial parameter with respect to accuracy, time and error involved in prediction of several diseases like, diabetes, cardiovascular disease and son. The objective is to propose a health care monitoring system which will improve the QoS in terms of clustering accuracy, clustering time and error. Specifically for IoT with sensors collecting patient's data, large amount of data is said to be created by sensors while extracting healthcare data. However, a machine learning technique can be utilized to learn the patient's data and process accordingly, however suffer from serious issues, henceforth affecting time, accuracy and error involved in disease prediction. In such circumstances to monitor the cardiovascular disease with cluster tendency aspect, a cluster threshold is utilized to map the cluster tendency with the actual cluster formation.

1.2 Our Contributions

Our contributions include the following:

1. An IoT health care monitor system has been proposed based on machine learning for cardiovascular disease prediction in case of large involvement of data (i.e., Big Data).
2. The patient's healthcare monitor based cardiovascular disease prediction model is proposed to advice the human being affected with disease regarding further treatment, to have control on certain aspects concerning disease and son. The healthcare monitoring method shows better accuracy than the existing methods.
3. A Theil–Sen Estimated Linear Regression Feature Selection model is first designed to select the relevant features or attributes for cardiovascular disease monitoring via multiple linear regressions and Theil–Sen Estimator function.
4. A Canopy Hopkins Statistic Clustering algorithm for healthcare monitoring is designed with the selected feature for detecting either the presence or absence of disease by employing appropriate learning parameters.
5. The proposed method has been experimented and validated with 70000 numbers of samples. The prediction accuracy, prediction time and error for different numbers of samples or patients while using the proposed method are determined and experimental analysis presents that the proposed method outperforms the state-of-the-art methods.

1.3 Organization of the work

The rest of the paper is organized as follows. Section 2 presents the related works. The proposed Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) for IoT-based healthcare monitoring method is described in Section 3. Section 4 analyses the performance of the proposed method followed by discussion in Section 5. Finally, the conclusion is presented in Section 6.

2. Related Works

The study of diseased patient health monitoring is crucial to improve one's quality of life. The only key for controlling and controlling millions of people's health would be big data and IoT owing to the reason that most of the countries are scarcity of medical experts.

An elaborative survey on Machine Learning-Based Big Data for IoT-Enabled healthcare monitoring system was investigated in [5]. Smart healthcare monitoring is mushrooming owing to the Internet of Things (IoT) with Big Data. The IoT along with deep learning in the healthcare mitigate diseases by transforming healthcare from face-to-face to telemedicine. In [6], a deep learning-based IoT enabled real-time health monitoring method was proposed. Also with the aid of cross validation test the method was found to be improved in terms of precision and recall.

Numerous researchers in the recent years have considered patients' health and bestowed an optimal and pertinent solution. With the employment of technologies such as Internet of Things and 5G, information can be transmitted between users in a more timely and secured manner. The IoT provides numerous advantages in the area of e-health.

In [7], a prioritization system was utilized with the purpose of prioritizing the sensitive information in IoT. Moreover, LSTM deep neural network was also applied for classifying and monitoring patients' condition in a remote manner that was contemplated as a paramount feature, therefore contributing to accuracy. Basic concepts and main elements of multimodal sensing information collection, optimization of AI-assisted telemedicine was also introduced in [8] for emotion classification prediction.

As far as smart healthcare systems are concerned, the patients' are monitored in a remote manner to put a full stop to the disease spread and bestow cost efficient treatment. The amalgamation of IoT-enabled healthcare monitoring and machine learning is contemplated as a perfect solution.

A blockchain-based secured healthcare monitoring systems for diabetes disease employing adaptive neuro fuzzy inference was proposed in [9], therefore contributing to prediction accuracy. A holistic comparative prognosis prediction machine learning methods using recursive feature elimination via support vector machine was presented in [10] for early mortality prediction. In [11], a novel and intelligent healthcare system based on IoT and machine learning was proposed with the objective of sensing and processing patient's data via medical decision support system. With this design both the cost and time involved in prediction were found to be efficient.

IoT has not only intensified the liberty but also manifold the potentiality of the human to interact with external environment. IoT, with assistance of futuristic materials and methods became a paramount contributor to communication globally.

In [12], the current work bestows a holistic source of information concerning the numerous areas of application of HIoT aspiring to assist subsequent researchers, who have the intrigue to work and make evolutions in the area to acquire perception into the subject matter. Machine learning based glucose prediction employing physical activity monitoring data was proposed in [13]. Also by employing spearman correlation coefficient, the prediction accuracy was improved to a greater extent.

Despite big data analysis and machine learning being significantly researched, there is a dearth of research that solely concentrate on the advancement of machine learning based methods for big data analysis in the IoT healthcare. In [14], a comprehensive review on the machine learning application for big data analysis in the healthcare sector was investigated. Moreover, the advantages and drawbacks of existing methods in addition to numerous research challenges were also highlighted. A systematic review for healthcare using machine learning was designed in [15].

Big Data initiated with a blend of organized and un-organized layouts are tremendously large in dimensions and diversified. The Big Data is growing fiercely and has hitherto been lengthened to an incredible scale. This could be owing to the evolution of internet, IoT and social media. The prospects from smart clothing with a concentration on application to healthcare were investigated in [16]. In [17] a study on technology integrated health management was analyzed for ceaseless monitoring of patient with dementia. With different temporal granularity recognition accuracy was improved to a greater extent.

In [18], a new method of wearable sensor device for acquiring real time athlete data using IoT for monitoring electrocardiogram (ECG) patterns in addition to the body acceleration using smart phone was proposed. Also with the obtained data classification using Radial-basis Function Network and Levenberg-Marquardt with Probabilistic Neural Network was performed for health monitoring. Some of the issues and gaps in IoT for healthcare monitoring were analyzed in [19]. A comprehensive review using deep learning for IoT in healthcare was investigated in [20].

Motivated by the above research works and to fill the gaps in this work, a method called, Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC), for IoT-based healthcare monitoring is proposed. The elaborate description of the method is discussed in the forthcoming sections.

3. Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering for IoT-based healthcare monitoring

Numerous researchers have in the recent few years contemplated patients' health and bestowed a perfect and suitable solution. With the emergence of technologies like, Internet of Things and Big data, information can be interchanged swiftly and in a more secure manner. The Internet of things (IoT) with Big Data provides numerous opportunities in the field of e-health care monitoring. Using IoT with Big Data in this process can notably enhance the patient monitoring. Hence, it becomes paramount to provide a useful method in the medical industry to monitor the patients' status employing associated sensors.

In this section we plan to develop a method called, Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC), for IoT-based healthcare monitoring with higher accuracy and lesser time consumption. The proposed TSLR-

CHSC method is split into three sections. Figure 1 shows the structure of the proposed Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) method.

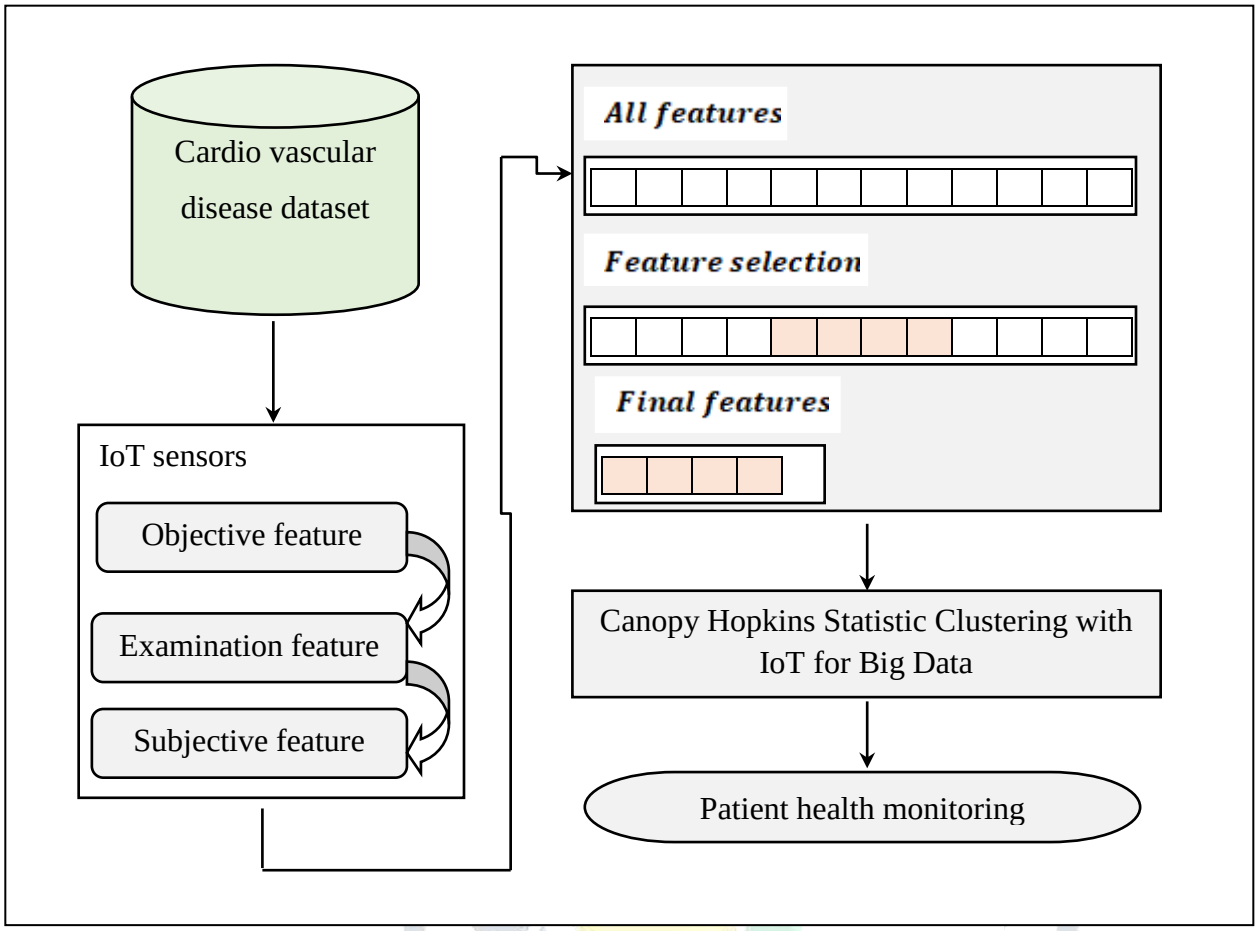

Figure 1 Structure of Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) method

As shown in the above figure, the TSLR-CHSC method is split into three sections, namely, data collection, feature selection and clustering. In the data collection section, IoT patient data is collected from the cardio vascular disease dataset. Second, to select the relevant features in a computationally efficient manner, Theil–Sen Estimated Linear Regression feature selection model is employed. Finally, with the computationally efficient selected features, a clustering model called, Canopy Hopkins Statistic Clustering is carried out to diagnose disease in a significant manner concentrating on the error aspect. In this manner, an efficient diseased patient health monitoring is carried out in a timely manner with improved accuracy and error rate.

3.1 Data Collection

To start with, in the proposed TSLR-CHSC method, IoT patient data is collected from the database. The objective of proposed TSLR-CHSC method is to perform a data analysis of cardiovascular patients employing the data acquired from kaggler. The dataset sourced from Kaggle at https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. The dataset used in this TSLR-CHSC method is obtained from kaggle as mentioned earlier. This data has 70000 observations with 12 descriptive features and 1 target excluding the ID column. The features or the attributes in the dataset are provided in the table 1 as given below.

Table 1 Cardiovascular disease dataset details

S. No	Features or attributes	Description
1	AGE	Integer (years of age)
2	HEIGHT	Integer (cm)
3	WEIGHT	Integer (kg)
4	GENDER	Categorical (1: female; 2:male)
5	AP_HIGH	Systolic blood pressure (integer)
6	AP_LOW	Diastolic blood pressure (integer)
7	CHOLESTEROL	Categorical (1: normal; 2: above normal; 3: well above normal)
8	GLUCOSE	Categorical (1: normal; 2: above normal; 3: well above

		normal)
9	SMOKE	Categorical (0:no; 1:yes)
10	ALOCOHOL	Categorical (0:no; 1:yes)
11	PHYSICAL ACTIVITY	Categorical (0:no; 1:yes)
Target variable		
12	CARDIO_DISEASE	Categorical (0:no; 1:yes)

As given in the above table, three distinct types of input features are present. They are objective feature (i.e., factual information), examination feature (i.e., results of medical examination) and subjective feature (i.e., information given by patient) respectively. All of the dataset values were acquired from the sensors and were collected at the moment of medical examination.

3.2 Theil–Sen Estimated Linear Regression Feature Selection model

In this section with the collected IoT patient data from the database, with the objective of selecting the relevant features with minimum time and maximum accuracy, Theil–Sen Estimated Linear Regression Feature Selective Process. Bayesian linear regression being a linear regression pattern robustly fit a line to sample points in the plane by selecting the median of slopes of all lines by means of pairs of point data for performing efficient diseased patient health monitoring. With this the relevant features are selected with higher accuracy and lesser time consumption. Figure 2 shows the structure of Theil–Sen Estimated Linear Regression Feature Selection model.

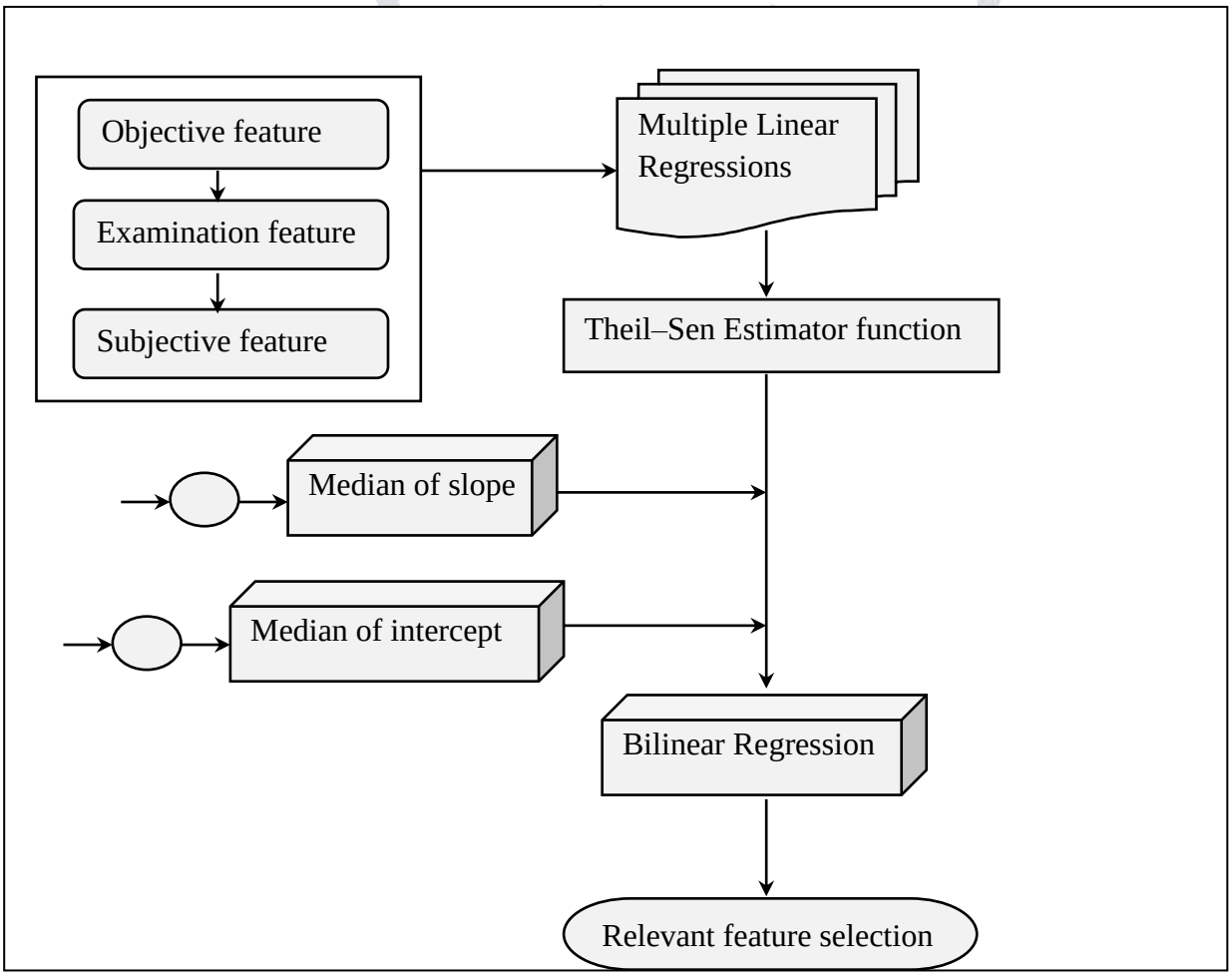

Fig. 2 Structure of Theil–Sen Estimated Linear Regression Feature Selection model

As shown in the above figure, with the three distinct features acquired from the sensors for analyzing healthcare concerning cardiovascular disease, first let us apply the Theil–Sen Estimator function via multiple linear regressions as given below.

$$Y_i = \alpha + F_i^T \beta + \epsilon_i \quad (1)$$

From the above equation (1), 'α' and 'β' symbolizes the intercept and slope of the corresponding feature 'F' sample points in the plane with a significant arbitrary errors 'ϵ_i'. Then, the slope 'β' of the corresponding feature 'F' sample points in the plane with two distinct points '$(F_i, Y_i), (F_j, Y_j), and\ (F_i \neq F_j)$' is required to find the association between features. So the slope of the corresponding feature sample points 'F' in the plane is mathematically formulated as given below.

$$\beta(q_{ij}) = \frac{Y_i - Y_j}{F_i - F_j} \quad (2)$$

With the above slope (2), the sum of squares of the corresponding feature sample points 'F' in the plane is formulated to find the association between features, 'F_i' and 'F_j', as given below.

$$(Y_i - \alpha - \beta F_i)^2 + (Y_j - \alpha - \beta F_j)^2 \quad (3)$$

$$where Y_i - \alpha - \beta F_i = 0\ and\ Y_j - \alpha - \beta F_j = 0$$

With the above sum of squares of residuals for two distinct points (3), the least square estimator of the corresponding feature sample points 'F' to find the association between features, 'F_i' and 'F_j' and discard irrelevant features present in the plane is given below.

$$p_{ij} = Y_i - q_{ij} F_i; q_{ij} = \frac{Y_i - Y_j}{F_i - F_j} \quad (4)$$

The slope 'β_i', is then the median of the least square estimators that is mathematically formulated as given below.

$$\beta_i = median\ \{q\}; q = \left\{ q_{ij} | q_{ij} = \frac{Y_j - Y_i}{F_j - F_i} \right\} \quad (5)$$

In a similar manner, the intercept 'α_i' is then the median of the least square estimators mathematically formulated as given below

$$\alpha_i = median\ \{p\}; p = \left\{ p_{ij} | p_{ij} = \frac{(Y_j F_i - Y_i F_j)}{(F_i - F_j)} \right\} \quad (6)$$

Finally, the Bayesian linear regression with respect to the slope and intercept to find the relevant feature is formulated as given below.

$$Prob\ (\beta|Y,F) = \frac{Prob\ (Y|\beta,F) * Prob\ (\beta|F)}{Prob\ (Y|F)} \quad (7)$$

$$Prob\ (\alpha|Y,F) = \frac{Prob\ (Y|\alpha,F) * Prob\ (\alpha|F)}{Prob\ (Y|F)} \quad (8)$$

$$FS = Prob\ (\beta|Y,F) \cup Prob\ (\alpha|Y,F) \quad (9)$$

From the above equations (7) and (8), '$Prob\ (\beta|Y,F)$', '$Prob\ (\alpha|Y,F)$' refers to the posterior probability distribution of the model parameters given the inputs and outputs. This is equal to the likelihood of the data, '$Prob\ (Y|\beta,F)$', '$Prob\ (Y|\alpha,F)$' multiplied by the prior probability of the parameters '$Prob\ (\beta|F)$', '$Prob\ (\alpha|F)$' and divided by normalization constant '$Prob\ (Y|F)$' respectively. The pseudo code representation of Theil–Sen Estimated Linear Regression Feature Selection is given below.

Input: Dataset 'DS', Features 'F'
Output: Computationally efficient and accurate feature selection 'FS'
Step 1: **Begin**

Step 2: **For** each Dataset 'DS' with Features 'F'
Step 3: Formulate multiple linear regressions as in equation (1)
Step 4: Estimate slope of the corresponding feature sample points 'F' in the plane as in equation (2)
Step 5: Evaluate sum of squares of residuals as in equation (3)
Step 6: Evaluate least square estimators as in equation (4)
Step 7: Estimate intercept of the corresponding feature sample points 'F' in the plane as in equation (5)
Step 8: Evaluate intercept median of the least square estimators as in equation (6)
Step 9: Estimate Bayesian linear regression with respect to the slope and intercept as in equations (7) and (8)
Step 10: **Return** relevant features 'FS' as in equation (9)
Step 11: **End for**
Step 12: **End**

Algorithm 1 Theil–Sen Estimated Linear Regression Feature Selection

As given in the above Theil–Sen Estimated Linear Regression Feature Selection with the distinct features collected from the corresponding sensors, the objective remains in selecting the relevant feature in a timely and accurate manner. With this objective, first, multiple linear regressions are formulated for distinct features present in the dataset. Followed by which the median of the slope and intercept fitting a line (i.e., to analyze the cardio vascular disease) to the sample points (i.e., values of the features) in the plane (i.e., for the overall dataset) is estimated. Finally, the best optimal features are obtained by employing Bayesian linear regression, therefore contributing to both time and accuracy.

3.3 Canopy Hopkins Statistic Clustering for healthcare monitoring

With the relevant feature selected in minimum time and maximum accuracy, in this section, Canopy Hopkins Statistic Clustering process is carried out to diagnose disease in an efficient manner. Canopy clustering being an unsupervised pre-clustering process, initializes two clusters, one for normal condition and the other for abnormal condition. Followed by which, finally, with the nature of data being Big Data, Hopkins Statistic Analysis is carried out for determining cluster tendency to what degree clusters exist in data to be clustered. In this manner, an efficient diseased patient health monitoring is carried out with minimal error rate. Fig 3 shows the structure of Canopy Hopkins Statistic Clustering for healthcare monitoring.

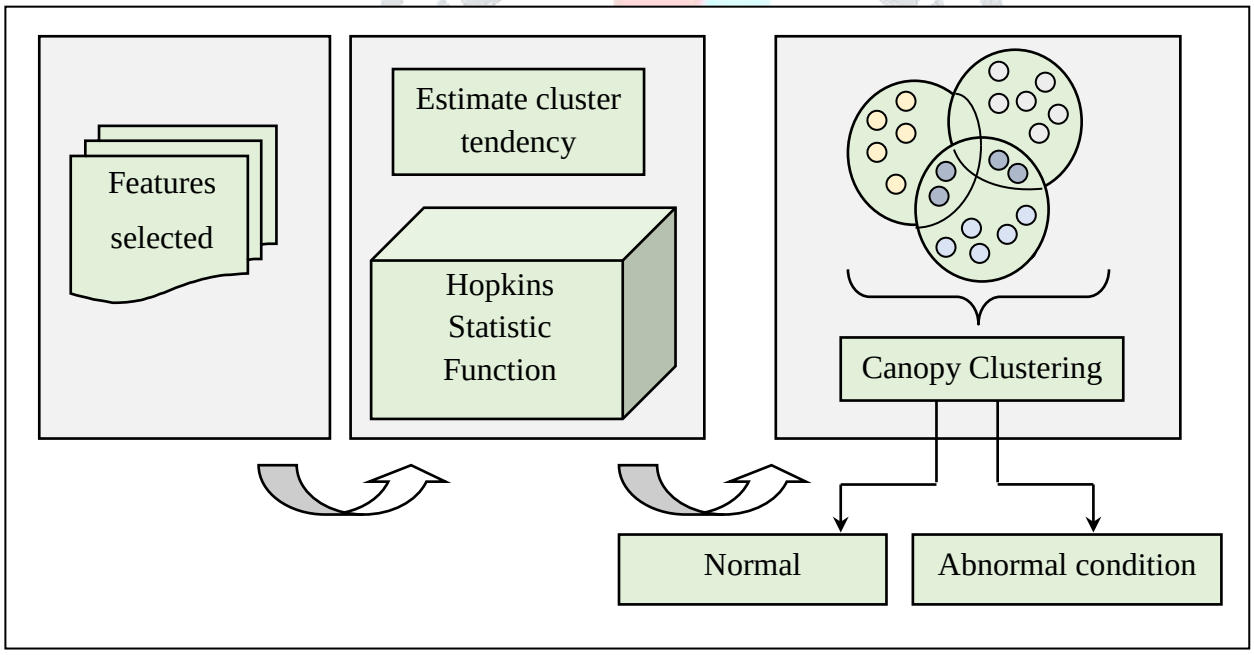

Fig 3 Structure of Canopy Hopkins Statistic Clustering for healthcare monitoring

As shown in the above figure, before proceeding with the actual clustering process for obtaining two distinct classes (normal condition class and abnormal condition class), first, the cluster tendency has to be evaluated to analyze whether the features selected from the cardiovascular disease dataset possess meaningful clusters or not. This is owing to the reason that as the nature of data is Big Data without the proper cluster tendency analysis, a lot of time and inaccurate grouping of objects into clusters are formed, resulting in the error. To address this issue, in our work, Hopkins Statistic function is initially derived with respect to the selected features. For example let us consider the features selected 'FS' from the actual dataset 'D'. Then, the 'n'

points are sampled uniformly '$(a_1, a_2, \dots, a_n)$' from the features selected 'FS'. For each point 'a_i' the nearest neighbor 'a_j' is identified and evaluates the distance between two points as given below.

$$g_i = Dist(a_i, a_j) \tag{10}$$

Followed by which in the second step, simulated features in a random fashion '$random_{FS}$' with 'n' points '$(b_1, b_2, \dots, b_n)$' are obtained. Then, for each point 'b_i' the nearest neighbor 'b_j' is identified and evaluates the distance between two points as given below

$$h_i = Dist(b_i, b_j) \tag{11}$$

Finally, the Hopkins Statistic function with the average nearest neighbor distance in the random simulated set divided by the sum of the average nearest neighbor distance is mathematically formulated as given below.

$$H = \frac{\sum_{i=1}^{n} h_i}{\sum_{i=1}^{n} g_i + \sum_{i=1}^{n} h_i} \tag{12}$$

From the above equation (12), higher value of 'H' indicates better clustering tendency than with the lower value. Accordingly, Canopy Clustering process is applied for obtaining or classifying into two different classes. For each sample point, it is assigned to the new canopy it its distance to the first point of canopy is less than the threshold 'Th_1'. On the other hand, if the distance to the second point of canopy is greater than the threshold 'Th_2', it is removed. The steps are repeated until there are no more data points in the set to cluster. The pseudo code representation of Canopy Hopkins Statistic Clustering for healthcare monitoring is given below.

Input: Dataset 'DS', Thresholds 'Th_1', 'Th_2'

Output: Error minimized cluster forming accurate disease diagnosis

Step 1: **Initialize** features selected 'FS', canopy 'C'
Step 2: **Begin**
Step 3: **For** Dataset 'DS' with 'n' points that are sampled uniformly between 'a_i' and 'a_j'
Step 4: Evaluate distance between two points as in equation (10)
Step 5: **If** distance between two points $\leq$ 'Th_1'
Sep 6: **Then** '$C \rightarrow C \cup g_i$'
Step 7: **Go to step** 18
Step 8: **End if**
Step 9: **End for**
Step 10: **For** each simulated features randomly '$random_{FS}$' with 'n' points between 'b_i' and 'b_j'
Step 11: Evaluate the distance between two points as in equation (11)
Step 12: **If** distance between two points $\leq$ 'Th_2' remove (a_i, a_j)
Step 13: **Then** remove 'h_i'
Step 14: R**eturn** '$abnormal\ condition$'
Step 15: **Go to step** 4
Step 16: **End if**
Step 17: **End for**
Step 18: Estimate Hopkins Statistic function as in equation (12)
Step 19: **Return** '$normal\ condition$'
Step 20: **End**

Algorithm 2 Canopy Hopkins Statistic Clustering

As given in the above Canopy Hopkins Statistic Clustering, for each Dataset with 'n' points that are sampled uniformly between two points are obtained and accordingly the distance measure is obtained. Followed by which conditional checking is made between the sampled distance and the threshold and upon satisfaction of the condition Hopkins Statistic function the cluster is assigned as normal. On the other hand, conditional checking is made between the random distance and the threshold and upon satisfaction of the condition Hopkins Statistic function the cluster is assigned as abnormal condition. The process is repeated until no more data points are left. In this manner, as only after determination of the cluster tendency, clusters are formed the error rate involved during the clustering process and subsequently, disease diagnosis rate is improved significantly.

4. Experimental setup

The proposed method Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) for IoT-based healthcare monitoring was implemented using the Python. The performance of the proposed method is evaluated using the benchmark dataset cardiovascular disease dataset obtained from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. This dataset possess information acquired from 70000 records of patients' data at the moment of medical examination. Experimental evaluation were carried out on certain parameters like, clustering accuracy, clustering time and error rate for varied sample patient data collected at different time intervals. To ensure a fair comparison similar sample patient data were used for all the three methods, over 10 simulation runs.

4.1 Performance metrics

One of the paramount metrics of analysis for IoTs with Big Data related to healthcare domain is the rate of accuracy with which the clustering is made. This is due to the reason that the accuracy is directly linked to health sciences as diagnosis and treatment can also be made, therefore aggravating the disease further. The clustering accuracy is measured as given below.

$$CAcc = \sum_{i=1}^{n} \frac{PD_{CA}}{PD_i} * 100 \quad (13)$$

From the above equation (13), clustering accuracy '$CAcc$' is measured on the basis of the samples or the patients data involved in experimentation 'PD_i' and the patients data accurately clustered 'PD_{CA}'. It is measured in terms of percentage (%).The second factor required for healthcare monitoring of the cardiovascular disease influencing the patient is the clustering time. This is because of the reason that early the clustering made, faster can be the treatment given to the patient and therefore reducing the rate of aggravate. The clustering time is evaluated as given below.

$$CTime = \sum_{i=1}^{n} PD_i * Time\ [Clustering] \quad (14)$$

From the above equation (14), the clustering time '$CTime$' is measured on the basis of the patient data involved in clustering 'PD_i' and the time consumed in clustering '$Time\ [Clustering]$'. It is measured in terms of milliseconds (ms). Finally, the error rate one of the critical statistical measures formulated for healthcare monitoring is mathematically formulated as given below.

$$ErrRate = \sum_{i=1}^{n} \frac{PD_{WC}}{PD_i} * 100 \quad (15)$$

From the above equation (15), the error rate '$ErrRate$' is measured based on the patient data involved in the simulation process 'PD_i' and the patient data wrongly clustered 'PD_{WC}'. It is measured in terms of percentage (%).

4.2 Results and Discussions

In this section, the results analysis for three different parameters, clustering accuracy, clustering time and error rate for healthcare monitoring with IoT and Big Data is discussed with respect to cardiovascular disease. Comparison is made with the proposed called Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) and existing methods, machine learning-based healthcare [1] and PATH2iot [2] using cardiovascular disease dataset from Kaggle.

4.2.1 Performance analysis of clustering accuracy

Several performance metrics were utilized to estimate the efficiency of the proposed TSLR-CHSC method. Clustering accuracy represents the overall clustering potentiality of the proposed machine learning method. Table 2 given below shows the results of the performance evaluation of the clustering accuracy using the three methods, TSLR-CHSC, machine learning-based healthcare [1] and PATH2iot [2] respectively.

Table 2 Performance evaluation of clustering accuracy

Number of patient data	Clustering accuracy (%)		
	TSLR-CHSC	**machine learning-based healthcare**	**PATH2iot**
7000	97.21	96.5	95.5

14000	95.15	94.35	93.15
21000	94	92	90.55
28000	93.25	90.35	89.25
35000	93	88.15	87.35
42000	92.85	87.35	85
49000	92.55	86.85	84.35
56000	92	85.35	83.15
63000	90	85.15	82
70000	88.35	85	82

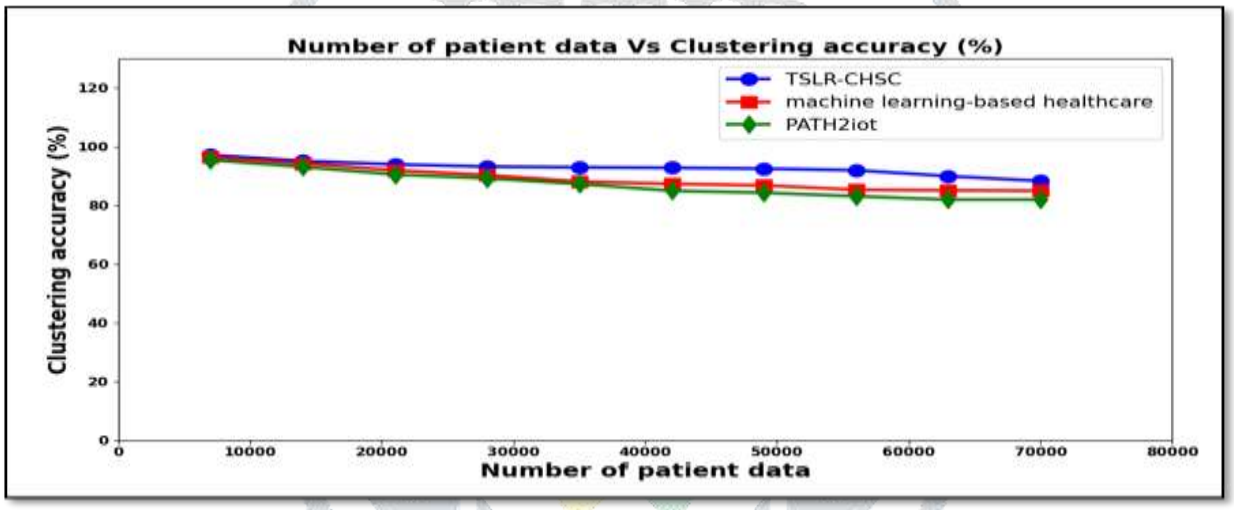

Fig. 4 Graphical representation of clustering accuracy

Fig. 4 given above illustrates the clustering accuracy with respect to 70000 distinct numbers of patient data. From the above figure, an inverse proportionate between number of patient data and clustering accuracy is found. To be more specific, increasing the number of patient data results in a proportionate decrease in the clustering accuracy. Let us consider a scenario with 7000 number of patient data, the patient data accurately clustered using TSLR-CHSC was found to be 6805, 6755 patient data accurately clustered using [1] and 6685 patient data accurately clustered using [2]. With this the overall clustering accuracy using the three methods were found to be 97.21%, 96.5% and 95.5% respectively. From this the clustering accuracy was found to be better using TSLR-CHSC method when compared to [1] and [2]. The reasons behind the improvement were owing to the application of Theil–Sen Estimated Linear Regression Feature Selection model. By applying this model, only relevant features involved in clustering process were selected by means of multiple linear regressions. This was again attained by splitting the features into slope and intercept that in turn selected the relevant features, therefore accurately clustering for further healthcare monitoring of cardiovascular disease using TSLR-CHSC method by 4% compared to [1] and 7% compared to [2].

4.2.2 Performance analysis of clustering time

In this section clustering time involved in healthcare monitoring for cardiovascular diseases using the proposed TSLR-CHSC is presented. Table 3 given below shows the results of the performance evaluation of the clustering time using the three methods, TSLR-CHSC, machine learning-based healthcare [1] and PATH2iot [2] respectively.

Table 3 Performance evaluation of clustering time

Number of patient data	Clustering time (ms)		
	TSLR-CHSC	**machine learning-based healthcare**	**PATH2iot**
7000	1085	1435	1995
14000	1135	1625	2125
21000	1245	1815	2345
28000	1355	2035	2555
35000	1515	2145	2815
42000	1735	2235	3015
49000	1955	2455	3235
56000	2055	2755	3515
63000	2145	3055	3815
70000	2355	3135	4015

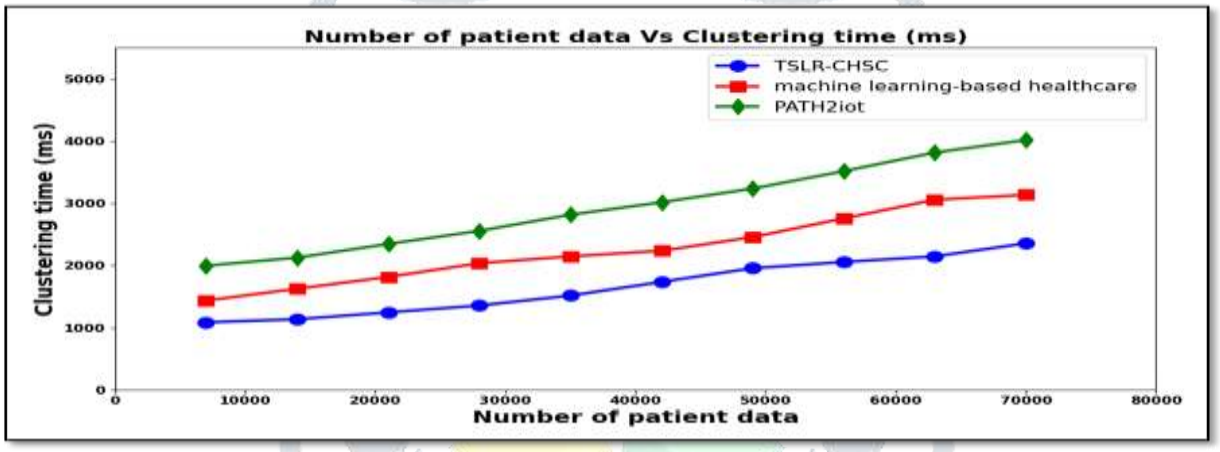

Fig 5 Graphical representation of clustering time

Fig 5 given above shows the clustering time for 70000 different numbers of patients' data acquired as input samples. From the figure it is observed that the clustering time is found to be directly proportional to number of patient data provided as input. To be more specific by increasing the number of patient data results in an increase in the testing made and also an increase in the patient's data collected as samples and obviously the clustering time was also found to be increased. With '7000' numbers of patients' data considered for experimentation and the time involved in clustering single sample being '$0.155ms$', time involved in clustering single sample being '$0.205ms$' using [1] and time involved in clustering single sample being '$0.285ms$' using [2], the clustering time was observed to be 1085ms, 1435ms and 1995 ms using the three methods. From the results it is inferred that the clustering time was found to be comparatively lesser using TSLR-CHSC when compared to [1] and [2]. From the results it is hypothesized that the clustering time using TSLR-CHSC is comparatively lesser than [1] and [2]. The reason behind the improvement is due to the incorporation of Bayesian linear regression with respect to the slope and intercept to find the relevant feature. With this relevant feature for clustering was obtained, therefore reducing the clustering time involved in healthcare monitoring with IoT and Big Data, therefore reducing the clustering using TSLR-CHSC by 27% compared to [1] and 44% compared to [2].

4.2.3 *Performance analysis of error rate*

Finally, the error rate involved in the clustering for healthcare diagnosis is measured. Table 4 given below shows the results of the performance evaluation of the error rate for three different methods, TSLR-CHSC, machine learning-based healthcare [1] and PATH2iot [2] respectively.

Table 4 Performance evaluation of error rate

Number of patient data	Error rate (%)		
	TSLR-CHSC	**machine learning-based healthcare**	**PATH2iot**
7000	2.64	3.64	5.07
14000	2.95	4.15	5.85
21000	3.05	4.55	6.15
28000	3.35	4.95	6.35
35000	3.95	5.05	6.85
42000	4.05	5.15	7
49000	4.15	5.55	7.15
56000	4.35	6	7.35
63000	4.55	6.35	7.85
70000	5	6.55	8

(Table 4 continued)

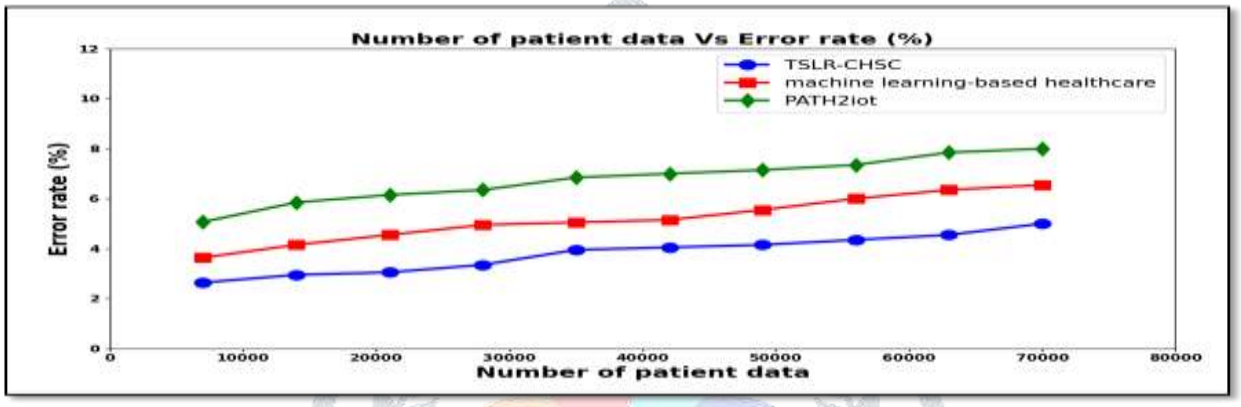

Fig. 6 Graphical representation of error rate

Figure 6 given above illustrates the error rate for three different methods. The error rate is one of the significant factors during clustering. This is due to the reason that not all the results are found to be accurately clustered into their appropriate classes and certain results are also wrongly clustered due to presence of noise in the selected features. From the figure a significant amount of rise is inferred with the increase in the number of patient data provided as samples in all the three methods from the cardiovascular disease dataset. However, with simulations conducted with 7000 samples, the error was observed to be 2.64% 3.64% and 5.07% using the proposed and the existing two methods respectively. With this measure, the error was reduced using TSLR-CHSC upon comparison with the two other existing methods. The reason behind the improvement was due to the application of Canopy Hopkins Statistic Clustering algorithm. By applying this algorithm, conscious healthcare monitoring was made by conditional checking made between the random distance and the threshold by employing Hopkins Statistic function. Accordingly two clusters, normal or abnormal were formed. With this, the error rate was significantly reduced in TSLR-CHSC by 27% compared to [1] and 44% compared to [2].

5. Conclusion

In this paper, machine learning based Theil Sen Linear Regression and Canopy Hopkins Statistic Clustering (TSLR-CHSC) method for IoT with Big Data is proposed to check vital features concerning cardiovascular disease and monitor the changes of persons via smart personnel care technologies. According to the theoretical model of the proposed method, two sections have been considered. These sections include obtaining the relevant feature or attribute for healthcare monitoring concerning cardiovascular disease employing Theil–Sen Estimator function and clustering via Canopy Hopkins Statistic Clustering algorithm. The proposed method was evaluated with different clustering methods. The applied clusters included machine learning and linear regression models. The experimental results revealed that the clustering algorithms using machine learning, called, TSLR-CHSC method performed well in terms of the clustering accuracy, clustering time and error rate. TSLR-CHSC method reached the highest performance for healthcare monitoring in our scenario with 97.21% accuracy, 1085ms time and error rate of 2.64%. High accuracy of TSLR-CHSC method in comparison with other applied clustering methods is a significant difference that makes it applicable in real-time healthcare monitoring.

References

[1] Alireza Souri, Marwan Yassin Ghafour, Aram Mahmood Ahmed, Fatemeh Safara, Ali Yamini and Mahdi Hoseyninezhad, “A new machine learning-based healthcare monitoring model for student’s condition diagnosis in Internet of Things environment”, Soft Computing, Springer, Volume 24, 2020, Pages 17111–17121 [machine learning-based healthcare]

[2] Devki Nandan Jha, Peter Michalak, Zhenyu Wen, Rajiv Ranjan, and Paul Watson "Multi-objective Deployment of Data Analysis Operations in Heterogeneous IoT Infrastructure", IEEE Transactions on Industrial Informatics, Volume 16, Issue 11, November 2020, Pages 7014 – 7024 [PATH2iot]

[3] R. Bharathi, T. Abirami, S. Dhanasekaran, Deepak Gupta, Ashish Khanna, Mohamed Elhoseny K. Shankar, "Energy Efficient Clustering with Disease Diagnosis Model for IoT based Sustainable Healthcare Systems", Sustainable Computing: Informatics and Systems, Volume 28, December 2020, Pages 1-28

[4] Hassan Harb, Ali Mansour, Abbass Nasser, Eduardo Motta Cruz and Isabel de la Torre Diez, "Sensor-Based Data Analytics for Patient Monitoring in Connected Healthcare Applications", IEEE Sensors Journal, Volume 21, Issue 2, January 2021, Pages 1-10

[5] Wei Li, Yuanbo Chai, Fazlullah Khan, Syed Rooh Ullah Jan, Sahil Verma, Varun G. Menon, Kavita, Xingwang Li, "A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System", Mobile Networks and Applications, Springer, Jan 2021

[6] Xingdong Wu, Chao Liu, Lijun Wang, Muhammad Bilal, "Internet of things-enabled real-time health monitoring system using deep learning", Neural Computing and Applications, Springer, Aug 2021

[7] Somayeh Iranpak, Asadollah Shahbahrami, Hassan Shakeri, "Remote patient monitoring and classifying using the internet of things platform combined with cloud computing", Journal of Big Data, Aug 2021

[8] Heng Yu, Zhiqing Zhour, "Optimization of IoT-Based Artificial Intelligence Assisted Telemedicine Health Analysis System", IEEE Access, Jun 2021

[9] P. G. Shynu, Varun G. Menon, R. Lakshmana Kumar, Seifedine Kadry, Yunyoung Nam, "Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing", IEEE Access, Mar 2021

[10] Mahdi Mahdavi, Hadi Choubdar, Erfan Zabeh, Michael Rieder, Safieddin Safavi-Naeini, Zsolt Jobbagy, Amirata Ghorbani, Atefeh Abedini, Arda Kiani, Vida Khanlarzadeh, Reza Lashgari, Ehsan Kamrani, "A machine learning based exploration of COVID-19 mortality risk", PLOS ONE | https://doi.org/10.1371/journal.pone.0252384 July 2, 2021

[11] Kashif Hameed, Imran Sarwar Bajwa, Shabana Ramzan, Waheed Anwar, Akmal Khan, "An Intelligent IoT Based Healthcare System Using Fuzzy Neural Networks", Scientific Programming, Hindawi, Dec 2020

[12] Bikash Pradhan, Saugat Bhattacharyya, Kunal Pal, "IoT-Based Applications in Healthcare Devices", Journal of Healthcare Engineering, Hindawi, Mar 2021

[13] William P. T. M. van Doorn, Yuri D. Foreman, Nicolaas C. Schaper, Hans H. C. M. Savelberg, Annemarie Koster, Carla J. H. van der Kallen, Anke Wesselius, Miranda T. Schram, Ronald M. A. Henry, Pieter C. Dagnelie, Bastiaan E. de Galan, Otto Bekers, Coen D. A. Stehouwer, Steven J. R. Meex, Martijn C. G. J. Brouwers, "Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study", PLOS ONE | https://doi.org/10.1371/journal.pone.0253125 June 24, 2021

[14] Wei Li, Yuanbo Chai, Fazlullah Khan, Syed Rooh Ullah Jan, Sahil Verma, Varun G. Menon, Kavita, Xingwang Li, "A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System", Mobile Networks and Applications, Springer, Jan 2021

[15] Aakansha Gupta, Rahul Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review", Journal of Biomedical Informatics, Elsevier, Jun 2020

[16] Mominul Ahsan, Siew Teay Hon, Alhussein Albarbar, "Development of Novel Big Data Analytics Framework for Smart Clothing", IEEE Access, Jan 2020

[17] Shirin Enshaeifar, Ahmed Zoha, Andreas Markides, Severin Skillman, Sahr Thomas Acton, Tarek Elsaleh, Masoud Hassanpour, Alireza Ahrabian, Mark Kenny, Stuart Klein, Helen Rostill, Ramin Nilforooshan, Payam Barnaghi, "Health management and pattern analysis of daily living activities of people with dementia using in-home sensors and machine learning techniques", PLOS ONE | https://doi.org/10.1371/journal.pone.0195605 May 3, 2018

[18] ZhonghuaWang, Zhonghe Gao, "Analysis of real-time heartbeat monitoring using wearable device Internet of Things system in sports environment", Computational Intelligence, Wiley, April 2020

[19] Jorge Calvillo-Arbizu, Isabel Román-Martínez, Javier Reina-Tosin, " Internet of things in health: Requirements, issues, and gaps", Computer Methods and Programs in Biomedicine, Elsevier, Jun 2021

[20] Hamidreza Bolhasani, Maryam Mohseni, Amir Masoud Rahmani, "Deep learning applications for IoT in health care: A systematic review", Informatics in Medicine Unlocked, Elsevier, Mar 2021

Authors Profile

I am V.Deepa M.Phil. and doing my Ph.D. Research Tiruppur Kumaran College for Women, P.G and Research Department of Computer Science, Tiruppur. Interested areas of Research are Data Mining, Image Processing and Big Data.

I am Dr K.Rajeswari M.Sc., M.Phil. Associate Professor, P.G and Research Department of Computer Science, Tiruppur Kumaran College for Women, Tiruppur. I am having more than 21 years of teaching experience and published more than 45 papers. Interested areas are Data Mining, Image Processing, Networking, and Big Data.