Mobile Health Application for Early Disease Outbreak-Period Detection

Preetika Rani
Dept. of Computer Sc. &
Engg., IIT Roorkee
preetika125@gmail.com

Vaskar Raychoudhury
Dept. of Computer Sc. &
Engg., IIT Roorkee
vaskar@ieee.org

Sandeep Singh Sandha
Dept. of Computer Sc. &
Engg., IIT Roorkee
sandha.iitr@gmail.com

Dhaval Patel
Dept. of Computer Sc. &
Engg., IIT Roorkee
patelfec@iitr.ac.in

Abstract: Mankind has experienced several deadly disease outbreaks, such as, cholera, plague, yellow fever, SARS, and dengue. Researchers need to study disease propagation data in order to understand patterns of disease outbreaks, their nature, symptoms, and ways of containment and cure. Though our healthcare establishments record and maintain patient information, they fail to detect a pandemic at an early stage due to the following challenges. Firstly, modern people are too busy to visit a doctor at the early stage of their symptoms which along with their high degree of mobility fuels the risk of contagion. Secondly, even for the recorded cases of a disease, quickly consolidating all local information to detect disease propagation over a large area is non-trivial using today's technology. Finally, all existing methods of outbreak detection identifies a single day of outbreak which is less realistic considering that outbreak happens over a period of time. In this paper, we introduce a wearable sensor based mobile application to capture early symptoms of a disease and to ensure faster consolidation of isolated cases over large areas. We then apply a purely novel technique based on discrepancy scores to detect disease outbreak-period. Experiments and prototypes show the usability and efficiency of our solution.

Keywords— Outbreak detection, outbreak period, discrepancy scores, smartphone, mobile healthcare, body-area sensor network

1. INTRODUCTION

Outbreaks of infectious diseases have wiped out vast population since the early stages of human civilization. Even with the miracles of medicine with which humans are armed today, we still find it difficult to prevent disease epidemics like, dengue or SARS. With the use of data management systems in hospitals, laboratories, emergency departments, and other medical facilities during recent decades, we have a huge dataset for us to learn patterns of outbreaks which can help to devise suitable methods for outbreak detection. Early detection of a disease outbreak is instrumental in its containment and finally finding a cure. However, that is surely non-trivial given the following challenges.

Firstly, due to their extremely busy schedule, today's people hardly find time to visit a doctor at the early stages of a disease. Moreover, we also make mistakes in identifying initial symptoms of critical illnesses if they are closely similar to common day-to-day ailments. So, in any way, our medical establishments are not even able to register diseases until the patients attain critical conditions.

Secondly, with the faster modes of communication, diseases spread even faster in modern times. E.g., after the first case of SARS was reported in Hong Kong in 2002, it spread into 37 countries of the world just within a few weeks. So, failure to detect early a disease outbreak may put the entire community into high risk.

Thirdly, all our medical databases are localized and specific to individual establishments. National boards collect data from different regions and then declare an outbreak when it is too late to recover or to administer early precautions. An automated central sever gathering data from all over the country and alerting about possible outbreaks in a timely manner can save lot of lives and resources.

Finally, the huge volume of medical records on disease outbreaks poses serious problems to the analysts while trying to fix a range of data on which one should focus. Different outbreak detection techniques only give the day when an outbreak has occurred. These techniques also have high rate of false positive cases. This happens due to factors like weather, season, weekends, etc. during which a particular disease peaks. However, they are not outbreaks. However, one can say that an outbreak cannot happen on a single day, but occurs over a period of time.

In order to address the aforementioned challenges, in this paper, we have proposed a system for early disease outbreak-period detection. Our research contribution is twofold. Initially, we devise a mobile application which makes use of several wearable health sensors and the user's smartphone to detect different vital and/or environmental signs (like, pulse rate, respiration rate, body temperature) on a continuous and real-time basis. The smartphone can analyze abnormal health conditions based on the symptoms sensed and can alert the user or the concerned authority in case of an emergency. Collected data is periodically sent to the backend server for long-term storage and processing. At the later stage we operate over the accumulated health data to detect outbreak period if there is any at all. We have used Discrepancy Score to find the disease outbreak period, by changing the definition of parameters in Discrepancy Score [1][2] formula.

Existing research works detect a disease outbreak by identifying a single day on which the reported number of cases (of a particular disease) crosses a pre-specified threshold over the baseline dataset. There are some rule-based anomaly pattern detection algorithms, like WSARE [4] which give us (on day basis) difference or strangeness about a particular day compared to the base line data. Early Abbreviation Reporting System (EARS) consist of three methods C1, C2, and C3 based on positive one-sided Shewhart control chart [5]. Shewhart control charts, Exponential Weighted Moving Average (EWMA), and Cumulative Sum (CUSUM) are

univariate statistical process control methods that detect whether a process is under control or an unexpected event has occurred [6]. They detect abnormal shifts in dataset over the baseline.

In summary, we make the following contributions in this paper.

- We have proposed a method to detect disease outbreak period, which, to the best of our knowledge, is purely novel.
- We have carried out extensive experiments to show the efficiency of our method. Our method minimizes the rate of false positives significantly.
- We have also developed a prototype mobile healthcare application to facilitate data collection from users on a real-time basis using body-area sensor networks. Also, our application helps to accumulate outbreak related data from a large population in a short while and to present it to the analyzers.

The rest of the paper is organized as follows. In Section 2 we introduce related research works in outbreak detection and mobile healthcare application. Section 3 describes our proposed method for outbreak-period detection. Our test results are reported in Section 4 and the prototype mobile application is introduced in Section 5. Finally, the paper concludes in Section 6 giving future directions for possible extensions.

2. Preliminaries

A time series $T = \{v[1], v[2],\ldots,v[n]\}$ with length $|T| = n$ is a sequence of regularly sampled real value observations where $v[i]$ is observation value at time i.

A period of time series, denoted as $T[i, j]$, is a subset of continuous observations starting at time i and ending at time j and has a length of $|T[i,j]| = (j-i+1)$.

In case of infectious disease outbreak data, we have two time series data of equal length. One time series having no outbreak is baseline dataset, while other time series we want to find outbreak period is test dataset.

An outbreak-period is defined as the occurrences of more cases of a particular disease than normally expected within a specific place and group of people, over a given period of time [7]. Point outbreak represents the day on which outbreak starts to occur. Outbreak period consists of continuous days in time series dataset, during which outbreak occurs. Outliers are some days in a year during which the number of patients are exceptionally high due to reasons other than a outbreak, e.g., hot weather.

3. Related work

We have summarized the related works on outbreak detection in Fig.1. Various methods like WSARE (What's Strange About Recent Event), EARS (Early Abbreviation Reporting System), statistical Methods, time period scan statistics, etc. are available for outbreak detection. These methods are different from each other based on the number of attributes required in dataset, past amount of dataset or baseline required to apply method, and the type of output given by the methods.

WSARE is a rule-based anomaly pattern detection algorithm Which works in two phases. In the first phase, the baseline that determines the "normal pattern" is estimated. Generally baseline for the current day includes 35, 42, 49, and 56 days prior to the current day. Second phase analyzes the recent patterns in a dataset and determines the patterns which are anomalous relative to the normal historical patterns. Each pattern is characterized by a rule (*flu=high* **OR** *reported_symptom=respiratory*) [4]. WSARE is very different from other traditional algorithms that focus on only one attribute of data, like *daily count of patient's records*. But there may be a case of insidious disease (i.e. disease which is proceeding in a slow manner but the effects are harmful), which affects only a small or specific group of people and hence, does not generate sufficient number of patient records, and so it cannot be detected by traditional algorithms. It does not tell us where in the data we have to focus or which time period is important from outbreak point-of-view. Data analysts have to analyse the whole database to find the outbreak because rules for each day may or may not be useful from outbreak point-of-view.
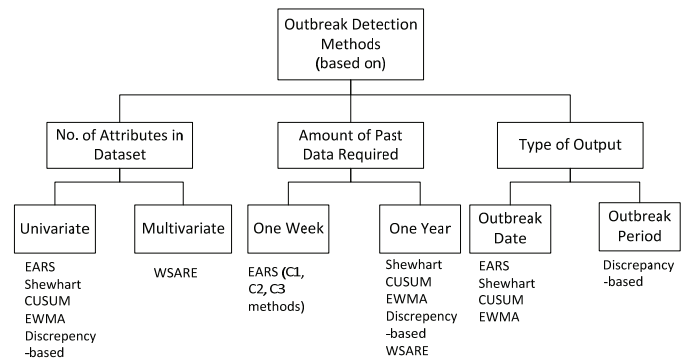

Fig.1: Classification of outbreak detection techniques

EARS are particularly used in detecting outbreaks by CDC (Centre for Disease Control, US). EARS contain three methods based on sensitivity C1-MILD, C2-MEDIUM, and C3-ULTRA methods. C1 is least sensitive while C3 is most sensitive to their respective baselines [5]. C1 method uses sample data from previous 7 days' data (with respect to the current day) and calculates moving average and standard deviation. C2 method is similar to C1 method except in taking the sample to calculate moving average and standard deviation. C2 method uses sample data from previous 7 days with 2-day-lag. C3 method use C_2 statistics value from day t to day (t-2) that is current day and previous two days, and calculate statistics C_3 for day t. C1 and C2 methods give signal when their statistics exceed the threshold value 3, while threshold value for C3 method is 2. These methods are very sensitive to past data of one week, so they give very high number of false positive outbreaks. We have summarized the EARS methods in Table 1.

A Shewhart control chart is a univariate statistical process control method that detects whether a process is under control or an unexpected event is occurring. If we have in-control

baseline, then from this we can compare current data and conclude whether this point is normal or an outbreak. It monitors the mean of processes, and if there is any variance which is above a threshold, then it raises an alarm. Shewhart charts are good in detecting large shifts from normal data but they are slow in detecting small shifts. They are also critical of the order (time) in which data is obtained [6].

Table 1: EARS Methods

C1- MILD	C2- MEDIUM	C3- ULTRA
$C_1(t) = \frac{Y(t) - \bar{Y}_1(t)}{S_1(t)}$	$C_2(t) = \frac{Y(t) - \bar{Y}_3(t)}{S_3(t)}$	$C_3(t) = \sum_{i=t}^{t-2} max[0, C_2(i) - 1]$
Where		
• $C_1(t), C_2(t), C_3(t)$ are C1, C2, C3 methods statistics • $Y(t)$ is total number of patient cases on day t • $\bar{Y}_1(t)$ and $S_1(t)$ is the mean and standard deviation of patient cases from day (t-1) to (t-7) • $\bar{Y}_3(t)$ and $S_3(t)$ is the mean and standard deviation of patient cases from day (t-3) to (t-9)		

EWMA is also two sided method i.e. it detects both increase and decrease in average of a distribution. As we are only interested in increase of number of disease incidence, one sided EWMA can be used to detecting increase. EWMA is good in detecting small shifts but not good in detecting large shifts in data as compared to Shewhart [6].

Table 2: Summary of Outbreak Detection Methods

Shewhart control charts	$St = \left| \frac{\bar{Y} - \hat{\mu}_0}{\hat{\sigma}_{\bar{Y}}} \right|$	where $\hat{\mu}_0$ is mean and $\hat{\sigma}_{\bar{Y}}$ is standard deviation of baseline data (365 days)
Exponential Weighted Moving Average (EWMA)	$E_t = \lambda \bar{Y}_t + (1 - \lambda)E_{t-1}$	where λ is weighing factor, $0 \geq \lambda \leq 1$, for detecting larger shifts if is preferable to take smaller value and to detect large shifts take larger values.
Cumulative Sum (CUSUM)	$C_t = max[0, C_{t-1} + L_t]$	where L_t is calculated by dividing residuals at time t by standard deviation of all 365 residuals

CUSUM is a successive test for an unknown distribution pattern (F_t) to check whether it is similar or dissimilar to a known pattern (F_0). As we are only interested in the rise in count of (disease) patients, one sided CUSUM is applicable. CUSUM charts are good in detecting small systematic effects in data, but cannot detect large shifts in data as compared to Shewhart. CUSUM find difficulty in finding special patterns. We modified the CUSUM, such that, when $C_t > h$, i.e., C_t crosses threshold (h) it does not reset the C_t value to zero and continuously compare C_t with threshold. CUSUM also tries to give the outbreak period but the time period is extended because of the cumulative increments on further days [6].

We have also studied several related research works in mobile healthcare application development. Zhu et al. [9] have developed smartphone based Android application for collection, transmission and analysis of pulse data. Lee et al. [10] have proposed a mobile healthcare system for elderly individuals which have been introduced in Taiwan hospitals. This system focuses on the automation of data storage and retrieval for long term tracking of patient's health status.

4. PROPOSED DISCREPANCY SCORES BASED METHOD

In this paper, our objective is to find out, if a pair of small segments (sporting number of patients with a particular disease), one each from baseline and test dataset plots, are having approximately same values. Here we assume that the two segments are of same size and belong to same calendar time period. If they are widely different, that may indicate the possibility of an outbreak. We are finding this difference with the help of Discrepancy Scores. Our method finds the discrepancy scores for different interval sizes of one, two, …, up to six days. In Fig.2 we have shown for one month, while in our experiment we have considered one year. As shown in Fig.2, we are calculating the discrepancy score for fixed-size interval (5 days) using four variables b, m, N_B, N_T. Here, b is the count of patients in a fixed interval (or a small segment) in the baseline dataset plot, m is the count of patients in a fixed interval (or a small segment) in the test dataset plot. Values of m and b are variable for different interval sizes, but the total count of patients in the entire baseline dataset (N_B) and the total count of patients in the entire test dataset (N_T) are constant [1][2].

$$\text{Discrepancy score(DS)} = m * \log\left(\frac{m}{b}\right) + (N_T - m) * \log\left(\frac{N_T - m}{N_B - b}\right)$$
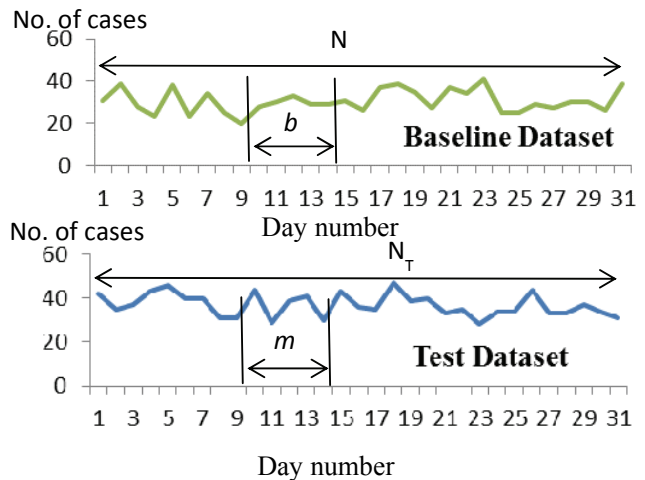
Fig.2: Values of parameters b, N$_B$, m and N$_T$

We have calculated the Discrepancy Scores (DS) for some intervals sizes, e.g., 1, 2, 3, 4.up to 10; 15, 20, 25, etc. (days). E.g., for a two-day interval size, values are calculated for day 1-2; day 2-3, day 3-4, and so on. We have finalized interval size of 6 days for detecting an outbreak period because

interval sizes of less than 6 days were generating large number of outbreak cases with short outbreak periods whereas, interval sizes of greater than 6 days were generating less number of outbreak cases with large outbreak periods. While the former case generates many false positives, the latter case is due to the overlapping of short outbreak periods found in interval sizes of less than 6, and in the worst case, it may identify the whole year as an outbreak period.

After finding the discrepancy scores for different intervals, we find the interval having discrepancy scores greater than the mean of all discrepancy scores. We are assuming that there is only one outbreak in a test year. So, if there are many intervals having their discrepancy scores greater than the mean discrepancy score, we choose the longest of such intervals as the outbreak period. Based on sensitivity of dataset, if the data is less sensitive then we find time period having Discrepancy Score greater than mean and three times standard deviation of Discrepancy Scores. If we find different periods having same longest length, then we select the period having greatest difference of discrepancy scores to mean discrepancy score.

5. PERFORMANCE ANALYSIS AND RESULTS

The dataset that we are using has been taken from autonlab [8]. The dataset is simulated for a city for two years; each patient in the city can do any one of the three actions, visit to emergency department, take a sick leave from school or work, or purchase medicine. Attributes of this dataset consists of *location*, *age*, *gender*, *flu level*, *day_of_week*, *weather*, *season*, *reported_symptom*, *drug*, and *date*. All of these are categorical attributes. Location is based on whether a case is from east, west, north or south. Flu level can be low, high, decline similarly other attributes have their respected values. The data of first year doesn't contain outbreak which is our baseline and second year has an outbreak of anthrax on specific date, which we know. We are using this dataset to find the period of outbreak or period which has abnormal number of patients; also this dataset is applied on various methods of outbreak detection.

Different methods were applied to dataset and the output was generated in different forms e.g., outbreak day, outbreak period, what strange about current day. The accuracy and false positive rate of different methods are different.

Table 3: Comparison between Existing Methods

| Outbreak Detection Methods		Number of point outbreak detected in a year (in days)	Average distance between two point outbreak (in days)
Early Aberration Reporting System (EARS)	C1 method	12	23.45
	C2 method	14	20.0
	C3 method	50	6.93
Shewhatrz Control chart		13	19.25
Exponential Weighted Moving Average (EWMA)		12	18.72
Cumulative Sum (CUSUM)		15	24.33

In Table 3 we provide a comparison based on point outbreak given by existing methods. There is only one outbreak in the whole year. C1 method of EARS shows 12 day on which there may be outbreak.

After calculating discrepancy scores for different intervals as described in Section 4, we check the time periods having discrepancy scores greater than mean. Among all such periods, we find the longest period that has greatest difference of mean and discrepancy scores (Table 4). We call that time period as "outbreak period". E.g., if we take the interval of size of one day, then the continuous days having the discrepancy scores much more than the mean is considered as outbreak period. Table 4 shows that for an interval of size one, the size of outbreak period is seven from 04-OCT (Day no. 277) to 11-OCT (Day no. 284). After taking union of all the intervals the length comes to be 45 days.

Table 4: Results of Discrepancy Scores for Different Period Sizes

Interval size (days)	Length of outbreak (days)	Starting date (Day no.)	Ending date (Day no.)
1	7	OCT-04-2003 (277)	OCT-11-2003 (284)
2	18	OCT-02-2003 (275)	OCT-20-2003 (293)
3	20	OCT-02-2003 (275)	OCT-22-2003 (295)
4	20	OCT-03-2003 (276)	OCT-23-2003 (296)
5	19	OCT-04-2003 (277)	OCT-23-2003 (296)
6	43	OCT-04-2003 (277)	NOV-16-2003 (320)

In Fig.3 the top most graph shows baseline Dataset (of a year) having no outbreak, then below that graph shown is test year we want to find outbreak. Bottom most graph show outbreak period given by Discrepancy score base method is shown with red colour (in square box). In the graphs in Fig. 3, X-axis shows the *number of cases registered*, Y-axis shows the *day number*.
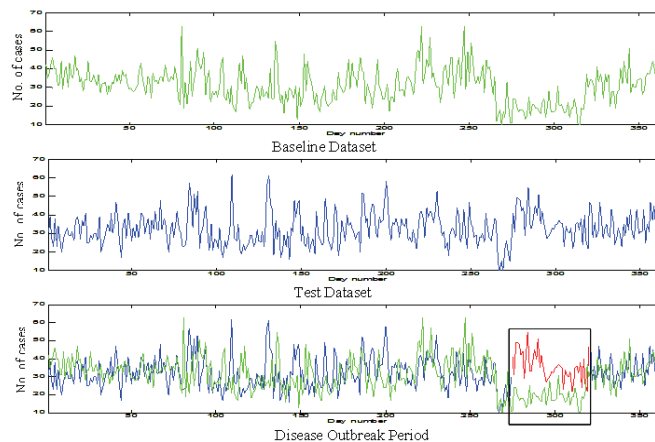
Fig.3: Results of applying our method on Baseline and Test Datasets

Fig.4 shows the results given by different methods (plotted on Y-axis) during the period of a year (365 days - plotted on X-axis). A peak represents an outbreak on a single day, while a continuous signals represents that outbreak took place over a period of time.
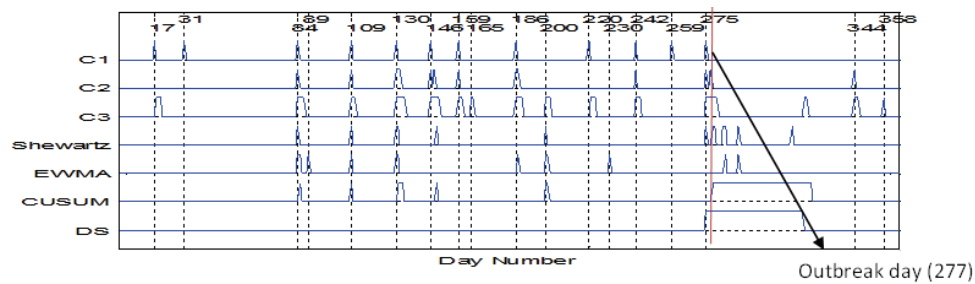
Fig.4: Comparison of Outbreak Detection Methods

Methods implemented by CDC that is EARS (C1, C2, and C3 method) give very frequent signals for outbreak so there are many false positive rates. While other methods gives very less signals as compared to them. More number of false positives will make it difficult for data analyst to focus on outbreak data, as there are so many signals given by these three methods. These three methods are very sensitive as they depend only on previous week data. These methods give signal when the outbreak period starts and as they are sensitive to past one week data, also they can't give signal again as the baseline they are considering is already having high number of cases.

Shewhart method give very frequent signals during outbreak but it also gives the false signals. The density of peaks is higher during outbreak period as compare to non-outbreak period. EWMA also gave false positive signal but it is not good in detecting outbreak earlier, it gives signal later when an outbreak has already occurred. CUSUM is good in detecting outbreak period because it continuously gives the signal during outbreak period. Last the discrepancy based method gives the period we are interested in. It gives very less or no false signals so that the data analyst can easily be focus on outbreak data.

6. PROTOTYPE MOBILE HEALTHCARE APPLICATION

In order to do Outbreak-period detection we need health related data of a large population at a central point, for which we have developed prototype mobile healthcare application which we have used for collecting early symptoms of a possible disease outbreak over a large population spread across wide area. Our application is implemented using a four-layer approach as shown in Fig. 5.

Health Monitoring Sensors: We have used following types of sensors to collect health-related data of patients.

- Patient Position Sensor (Accelerometer),
- Glucometer (measures blood glucose level)
- Body Temperature Sensor
- Blood Pressure Sensor (Sphygmomanometer)
- Pulse-oximeter (measures pulse rate and blood Oxygen level)
- Respiration Sensor (measures respiration rate)
- Galvanic Skin Response Sensor (measures skin resistance and conductance)
- Electrocardiogram (ECG) Sensor

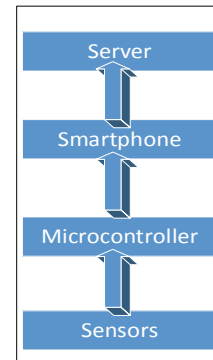- Electromyography (EMG) Sensor (measures electrical activity in skeletal muscles)

Fig.5: Four-layer Application Architecture

The readings collected by sensors are in analog format. They are voltage values normally varying from 0V to 5V. Sensors periodically update their values and send to microcontroller.

Microcontroller: We have used the Arduino MEGA ADK microcontroller board based on ATmega2560. Microcontroller, on receiving voltage values from sensors, it converts the analog voltage values to numbers on the scale of 0 to 100 and then them to smart phone.

Smartphone: Our application works on any Android based smartphone. For the testing purpose we have used two smartphones - a costly Samsung Galaxy S3 and a cheap Lava Iris 356. We have used two extreme capability phones to test our system. First is Samsung Galaxy S3 has Quad-core 1.4 GHz Cortex-A9 CPU and 1 GB RAM. Second is Lava Iris 356 has Dual-core 1 GHz CPU and 256 MB RAM. Our application works equally well in both the phones which goes to prove that it can be used by any average android phone with limited capabilities. Smartphone is the point where user can see his data in the form of numbers periodically and also can plot the data in real time. Smartphone also provide user the interactive interface where user can fill his basic details and then can send the data to online server.

Server: We used a PHP Web server as a back-end to store the health data captured in the smartphone. The server servers the two purposes, first it is global access of stored data and second it facilitates storage and analysis of health records for long term decision making like epidemic prediction.

Testing of the prototype mobile healthcare application:

The smartphone facilitates data visualization in terms of numerical values as well as graph plots. User can add his details, such as, name, age, weight and address, which are used to identify their personal health data in the Web server database. Application has *sent live* button, using which user can send his data to server when connected over internet. Fig. 6 shows the pictures of complete system with different sensors and smartphone.
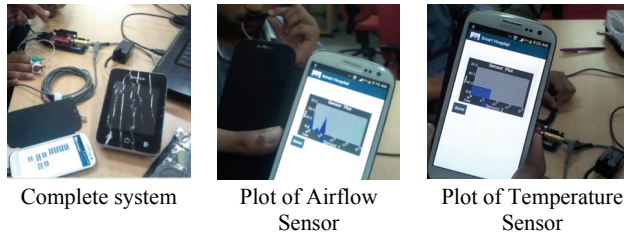

| Complete system | Plot of Airflow Sensor | Plot of Temperature Sensor |

Fig.6: Prototype Mobile Healthcare Application

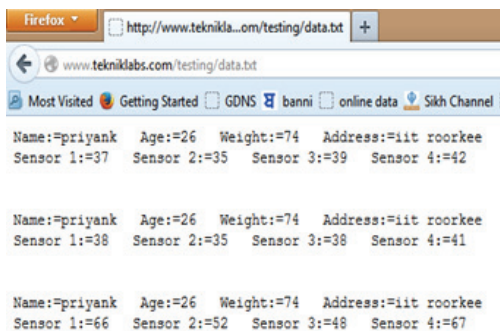**Fig.7: Screenshots of Mobile Healthcare Application**

Fig.8: Screenshots of Mobile Healthcare Server Interface

Fig. 7 show the screenshots of our application. We have identified different sensors as sensor 1, sensor 2, sensor 3, and sensor 4. Any health sensor can be connected to those ports and there can be a total of 16 sensors supported by a single micro-controller board. So, our application is generic in nature and can function with any type or number of sensors and any Android-based smartphone.

The interface of the server side is shown in Fig.8 with the data from 4 different sensors on a scale of 0-100 along with the user's personal details.

7. CONCLUSION

Outbreak cannot be detected at a point of time, because it happens over a period of time. Moreover, early outbreak detection is needed to minimize the sufferings and fatalities caused by an outbreak. In order to collect health data from a large population in a short time, we proposed a smartphone based mobile healthcare system which operates using several health sensors. Our system is cost-effective, portable, extendable, user friendly, and supports mobility. However, it generates large volume of data which poses a challenge while detecting outbreaks. In order to simplify the outbreak period detection, we have also proposed a novel method based on discrepancy scores. Our experiments and results go to prove that our system is efficient and minimizes the rate of false positives while compared with the existing outbreak detection approaches. This is because it ignores the outliers which happen for a short duration, as we assumed that the outbreak is happening for a period of time. This method can be generalized to any size of data. In future, we will try to reduce the number of attributes used for outbreak detection by categorizing them. We shall also try to combine our technique with existing outbreak detection techniques to achieve better results and to detect outbreak in lesser time.

ACKNOWLEDGMENTS

REFERENCES

[1] Chundi P. and Chen W., "Extracting hot spots of topics from time-stamped documents", in proc. of Data Knowl Eng. 2011 Jul; 70(7):642-660.

[2] Chen, W, and Chundi, P, "Trends analysis of topics based on temporal segmentation," Springer Berlin Heidelberg, 2009.

[3] "IBM What is big data?-Bringing big data to the enterprise", source (March, 2014): www.ibm.com.

[4] Wong W.K., et al., "What's strange about recent events (WSARE): An algorithm for the early detection of disease outbreaks." The Journal of Machine Learning Research 6 (2005): 1961-1998.

[5] Fricker Jr, R. D., et al. "Assessing the performance of the early aberration reporting system (EARS) Syndromic Surveillance Algorithms." *Statistics in Medicine* 27 (2008): 3407-3429.

[6] Fricker Jr., R.D., "Univariate Temporal Methods", in Introduction to Statistical Methods for Biosurveillance: With an Emphasis on Syndromic Surveillance. [Book] Cambridge University Press, 2013.

[7] "Epidemiology 101: Outbreak Investigation", source (March, 2014): http://infectiousdiseases.about.com/od/basics/a/outbreaks.htm

[8] "WSARE Dataset", source (March, 2014): www.autonlab.com

[9] Zhu, Tianyu, et al., "The exploitation and discussion of new mobile healthcare system model based on smart phone," In proc. of IEEE ICNSC'13 pp. 497-502.

[10] Lee, C-N, "A home care service platform for mobile healthcare," In proceedings of the IEEE International Conference on Machine Learning and Cybernetics, pp. 1927-1930, July 2012.