Intrusion Detection System Using Hierarchical GMM and Dimensionality Reduction

L. Maria Michael
Assistant Professor
Velammal Institute of Technology, Chennai

J. Indra Mercy
Student, M.E.
Saveetha Engineering College, Chennai

N.R. Rejin Paul
Assistant Professor
Velammal Institute of Technology, Chennai

ABSTRACT

The focus of this chapter is to provide the effective intrusion detection technique to protect Web server. The IDS protects an server from malicious attacks from the Internet if someone tries to break in through the firewall and tries to have access on any system in the trusted side and alerts the system administrator in case there is a breach in security. Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering the data. Intrusion detection can be divided into anomaly detection and misuse detection. Misuse detection model is to collect behavioral features of non-normal operation and establish related feature library. In the existing system of anomaly based Intrusion Detection System, the work is based on the number of attacks on the network and using decision tree analysis for rule matching and grading. We are proposing an IDS approach that will use signature based and anomaly based identification scheme. And we are also proposing the rule pruning scheme with GMM(Gaussian Mixture Model). It does facilitate efficient way of handling large amount of rules. And we are planned to compare the performance of the IDS on different models. The Dimension Reduction focuses on using information obtained KDD Cup 99 data set for the selection of attributes to identify the type of attacks. The dimensionality reduction is performed on 41 attributes to 14 and 7 attributes based on Best First Search method and then apply the two classifying Algorithms ID3 and J48 Keywords-Intrusion detection, reliable networks, malicious routers, internet dependability, tolerance.

Keywords

ID3, KDD, IDS, Dimensionality Reduction, NIDS.

1. INTRODUCTION

IDS is concerned with the detection of hostile actions. This network security tool uses either of two main techniques. The first one, anomaly detection, explores issues in intrusion detection associated with deviations from normal system or user behavior. The second employs signature detection to discriminate between anomaly or attack patterns (signatures) and known intrusion detection signatures. Both methods have their distinct advantages and disadvantages as well as suitable application areas of intrusion detection. Data are grouped using the rule pruning scheme with GMM (Gaussian Mixture Model). It does facilitate efficient way of handling large amount of rules.The issue of the web servers safety consists of two parts: One is the transmission security, including data on antieaves dropping and data integrity; the other is the web server side and client-side in itself. The former can be enhanced by a variety of security protocols. However, the latter need taking precautions by firewall and intrusion detection techniques. Generally speaking, the anti-attack of firewall is powerful, but not invulnerable. Intrusion detection techniques need to be applied to protect the web server because merely relying on the firewall is not enough.

Host intrusion detection refers to the class of intrusion detection systems that reside on and monitor an individual host machine. A network intrusion detection system monitors the packets that traverse a given network link. Network data has a variety of characteristics that are available for a NIDS to monitor: most operate by examining the IP and transport layer headers of individual packets, the content of these packets, or some combination thereof.

In this paper data mining classification algorithm is being used with the concept of Dimension Reduction. Dimension Reduction is applied using Best First Search which reduces the feature selection from 41 attributes to 14 and 7 potential attributes for classification. The proposed approach focuses on using information obtained KDD Cup 99 data set for the selection of attributes to identify the type of attack and then compares the performance of the ID3 with J48 by a randomly selected initial dataset with the reduced dimensionality. Furthermore, the results indicate that our approach provides more accurate results compared to the purely random one in a reasonable amount of time.

2. ANALYSIS TECHNIQUE

Misuse Detection: The essence of misuse detection centers around using an expert system to identify intrusions based on a predetermined knowledge base. As a result, misuse systems are capable of attaining high levels of accuracy in identifying even very subtle intrusions that are represented in their expert knowledge base. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks; as long as they have been Labeled appropriately. Their disadvantage is that they cannot detect unknown intrusions and they rely on signatures extracted by human experts. This method uses specifically known patterns of unauthorized behavior to predict and detect subsequent similar attempts. These specific patterns are called signatures.

The essence of misuse detection centers around using an expert system to identify intrusions based on a predetermined knowledge base. As a result, misuse systems are capable of attaining high levels of accuracy in identifying even very subtle intrusions that are represented in their expert knowledge base. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks; as long as they have been labeled appropriately. Their disadvantage is that they cannot detect unknown intrusions and they rely on signatures extracted by human experts. This method uses specifically known patterns of unauthorized behavior to predict and detect

subsequent similar attempts. These specific patterns are called signatures.

Anomaly Detection

Anomaly detection is concerned with identifying events that appear to be anomalous with respect to normal system behavior. Its Designed to uncover abnormal patterns of behavior, the IDS establishes a baseline of normal usage patterns, and anything that widely deviates from it gets flagged as a possible intrusion. Thus these techniques identify new types of intrusion as deviations from normal usage. It is an extremely powerful and novel tool but a potential drawback is the high false alarm rate, that is. previously unseen (yet legitimate) system behaviours may also be recognized as anomalies, and hence flagged as potential intrusions. If a user in the graphics department suddenly starts accessing accounting programs or compiling code, the system can properly alert its administrators.

In a network-based system, or NIDS, the every individual packet flowing through a network is analyzed. The NIDS can detect malicious packets that are designed to be overlooked by a firewall simplistic filtering rules. In a host-based system, the IDS examines at the activity on each individual computer or host.

A wide variety of techniques including neural networks, decision tree approach and hidden Markov models have been explored as different ways to cluster the data for rule creation. Each and every techniques has got its own pros and cons, Hidden markov model is slow, full search on a database of 400,000 sequences can take 15 hours. Decision tree approach is unstable to handle large volume of data,

Data Collection Issues: For accurate intrusion detection, we must have reliable and complete data about the target system's activities. Reliable data collection is a complex issue in itself. Most operating systems offer some form of auditing that provides an operations log for different users. These logs might be limited to the security-relevant events (such as failed login attempts) or they might offer a complete report on every system call invoked by every process. Similarly, routers and firewalls provide event logs for network activity. These logs might contain simple information, such as network connection openings and closings, or a complete record of every packet that appeared on the wire.

The amount of system activity information a system collects is a trade-off between overhead and effectiveness. A system that records every action in detail could have substantially degraded performance and require enormous disk storage. For example, collecting a complete log of a 100-Mbit Ethernet link's network packets could require hundreds of Gbytes per day.

3. OVERVIEW OF GAUSSIAN MIXTURE MODEL

Mixture models are a type of density model that comprise of a number of component functions, usually Gaussian. The distribution of feature vectors was extracted from packets in the network. A Gaussian Mixture Model GMM is used to construct a Bayesian classification procedure on the observations and leads to the system behavior model. Parameters of mixture model are used by the Expectation Maximization (EM) algorithm.

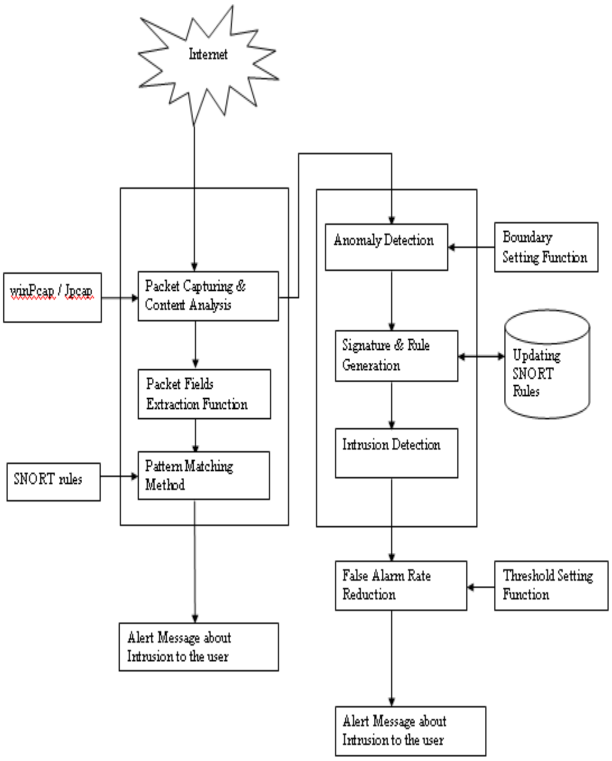

Internet
winPcap / Jpcap
Packet Capturing & Content Analysis
Packet Fields Extraction Function
SNORT rules
Pattern Matching Method
Alert Message about Intrusion to the user
Anomaly Detection
Boundary Setting Function
Signature & Rule Generation
Updating SNORT Rules
Intrusion Detection
False Alarm Rate Reduction
Threshold Setting Function
Alert Message about Intrusion to the user

Fig.1. 1 Overall system Architecture

Reducing The Data Features For Intrusion Detection Systems Using Gmm

The current intrusion detections aiming at the web server attack all adopt the rule-based method, like the famous intrusion detecting system Snort2.0, which detection rules are written after the features are refined from every intrusion behavior. Thus a rules library is formed. Then the captured data packets are matched the rules library respectively. If the match succeeds, the behavior is regarded as intrusion.

Since the amount of audit data that an IDS needs to verify is very huge even for a small network, rule matching is difficult even with computer assistance because extraneous features can make it harder to detect suspicious behavior patterns. Complex relationships exist between the features, which are difficult for humans to discover. IDS must group the amount of data to be processed. This is very important if real-time detection is desired. Reduction can occur in one of several ways. Data that is not considered useful can be filtered, leaving only the potentially interesting data. Data can be grouped or clustered to reveal hidden patterns; by storing the characteristics of the clusters instead of the data, overhead can be reduced. Finally, some data sources can be eliminated using feature selection

4. IMPROVING THE RULE MATCHING SPEED BY GMM

Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering. It deals with clustering problems: a model-based approach, which consists in using certain models for clusters and attempting to optimize the fit between the data. In practice, each cluster can be mathematically represented by a parametric distribution, like a Gaussian. The entire data set is therefore modeled by a mixture of these distributions. An individual distribution used to model a specific cluster is often referred to as a component distribution.

A mixture model with high likelihood tends to have the component distributions have high "peaks" and the mixture model "covers" the data well. Main advantages of model-based clustering:

well-studied statistical inference techniques

flexibility in choosing the component distribution;

obtain a density estimation for each cluster;

a "soft" classification is available.

Mixture of Gaussians

The most widely used clustering method of this kind is the one based on learning a mixture of Gaussians: we can actually consider clusters as Gaussian distributions centred on their barycentres, as we can see in this picture, where the grey circle represents the first variance of the distribution:

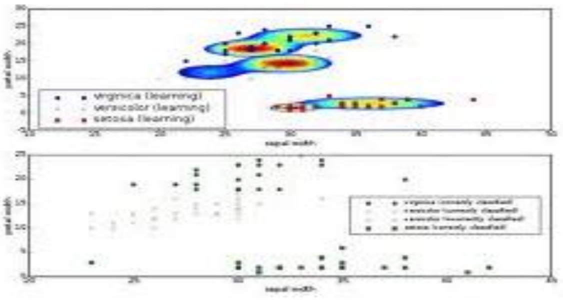


Fig 2. GMM Cluster

The algorithm first chooses the component (the Gaussian) at random with probability $P(\omega_i)$ and it it samples a point $N(\mu_i, \sigma^2 I)$.

Let's suppose to have:

x1, x2,..., xN

$P(\omega_1),\ldots,P(\omega_K),\sigma$

We can obtain the likelihood of the sample: $P(x \mid \omega_i, \mu_1, \mu_2, \ldots, \mu_K)$.

What we really want to maximise is $P(x \mid \mu_1, \mu_2, \ldots, \mu_K)$ (probability of a datum given the centres of the Gaussians).

$$P(x \mid \mu_i) = \sum_i P(\omega_i) P(x \mid \omega_i, \mu_1, \mu_2, \ldots, \mu_K)$$

is the base to write the likelihood function:

$$P(data \mid \mu_i) = \prod_{i=1}^{N} \sum_i P(\omega_i) P(x \mid \omega_i, \mu_1, \mu_2, \ldots, \mu_K)$$

Now we should maximise the likelihood function by calculating $\frac{\partial L}{\partial \mu_i} = 0$, but it would be too difficult. That's why we use a simplified algorithm called Expectation-Maximization

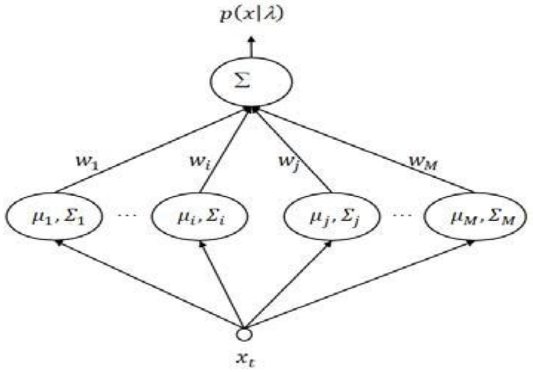


Fig. 3 Overview of the structure of GMM

In the Expectation Maximization algorithm no of mixtures decided beforehand, it updates the parameters of given k-component mixture with respect to the data set Xn = x1, ...xn such that likelihood of Xn is never smaller under new mixtures.

Estimates by iterating following equations for all components j €1, ..., k:

P(j|xi) = π j ϕ (xi; θ j)/fk(xi)

n
π j = ∑ P(j|xi)/n
i=1

n
μj = ∑ P(j|xi)xi/(n π j)
i=1

n
∑j = ∑ P(j|xi)(xi − μi)(xi − μi)T/(n π j)
i=1

Where θ is model with mean μ and covariace matrix ∑

π j: Mixing Weight

ϕ (x; θ j) : Mixture Component

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights

from all component densities.

Snort and Snort Rules

SNORT is one of the most popular NIDS. SNORT is Open Source, which means that the original program source code is available to anyone at no charge, and this has allowed many people to contribute to and analyse the programs construction. SNORT uses the most common open-source licence known as the GNU General Public License.

Rule generalization

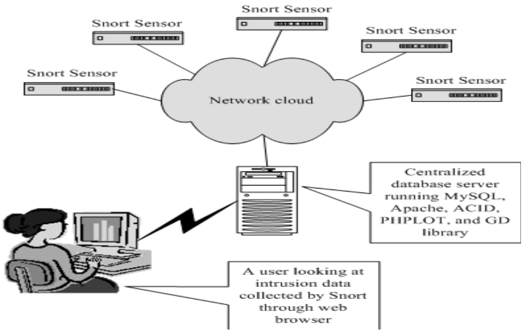
We propose to generate new rules by generalising SNORT rules. Given an Internet packet that contains a variation of a known attack, there should be some automated way to identify the packet as nearly matching a NIDS attack signature. If a particular statement has a set of conditions against it, an item may match some of the conditions.

Whereas Boolean logic would give the value false to the query 'does this item match the conditions', our logic could allow the item to match to a lesser extent rather than not at all. This principle can be applied when comparing an Internet packet against a set of conditions in a SNORT rule. Our hypothesis is that if all but one of the conditions are met, an alert with a lower priority can be issued against the Internet packet, as the packet may contain a variation of a known attack. In our implementation, generalisation in the case of matching network packets against rules, involves allowing a packet to generate an alert if:

The conditions in the rule do not all match, yet most of them do;

The only conditions that do not match exactly nearly match.

As an example, assume a certain rule states that an alert should be generated if a packet is a particular length, on a particular port and contained a certain bit pattern. Using our generalisation a packet matching those criteria, except perhaps on a different port, or with a slightly different bit pattern, would still count as matching, and a (modified) alert would be generated.


Snort Sensor
Snort Sensor
Snort Sensor
Snort Sensor
Snort Sensor
Network cloud
Centralized database server running MySQL, Apache, ACID, PHPLOT, and GD library
A user looking at intrusion data collected by Snort through web browser

Block diagram of a complete network intrusion detection system consisting of Snort, MySQL, Apache, ACID, PHP, GD Library and PHPLOT

5. PROPOSED SYSTEM

The hardware-implementable pattern matching algorithm for content filtering applications, which is scalable in terms of speed, the number of patterns and the pattern length. The algorithm is based on a memory efficient multihashing data structure called Bloom filter using embedded on-chip memory blocks in field programmable gate array/very large scale integration chips. In the proposed system we have detected anomalies from Internet connections, automated generation of rules and signatures for the anomalies Detection unknown or new attacks and reduced the false alarm rate. Models of network intrusion detection system are important component of the issue. Intrusion Detection System based on Back Propagation algorithm that can promptly detect attacks, no matter they are known or not. In this model, Back Propagation algorithm is used to learn about the normal users behavior and the abnormal users' behavior multigigabit per second speeds with a moderately small amount of embedded memory and a few mega bytes of external memory.

Dimensionality Reduction Algorithm

Dimension Reduction techniques are proposed as a data pre-processing step. This process identifies a suitable low-dimensional representation of original data. Reducing the dimensionality improves the computational efficiency and accuracy of the data analysis.

Steps :

□ Select the dataset.

□ Perform discretization for pre-processing the data.

□ Apply Best First Search algorithm to filter out redundant & super flows attributes.

□ Using the redundant attributes apply classification algorithm and compare their performance.

□ Identify the Best One.

The original dataset consist of 41 attributes and one class label. The following list out the attribute names

(i) 41 Attributes: duration, protocol type, service, Flag,, src_bytes, dst_bytes, land, wrong _ fragment, urgent,Hot,num_field_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creation, srv_count,.

serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_hosdst_same_srv_rate,dst_host_diff_srv_rate, dst_host_same _ src _ port _ rate, dst _ host _ srv _ diff _ host _ rate, dst _ host _serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate, dst _ host_srv_rerror_rate.

Using Best First Search method we obtained two set of reduced dimensionalities. 7 potential attributes and 14 potential attributes which are listed in the table 2 and 3 respectively.

(ii) 14 Attributes: duration, service, flag, src_bytes, dst_bytes, count, srv _ count, serror_rate, rerror_rate, dst _

host _ same _ srv _ count, dst_host_srv_rate, dst _ host _ rerror _ rate , dst _ host _ diff _ srv_byte, dst_host_

same _ src _port_rate.

(iii) 7 Attributes : Protocol Type, Service,Srcbytes, Dstbytes,count, diff_srv_rate, dest_host_srv_count,

Simulation Result

The Receiver Operating Characteristic (ROC) curve is usually used to measure the performance of the classification method. Here the ROC curve is a graphical plot of sensitivity, specificity for the attributes.

Table 1. Sensitivity, Specificity And Accuracy Based On 41 Attribute Feature Selections

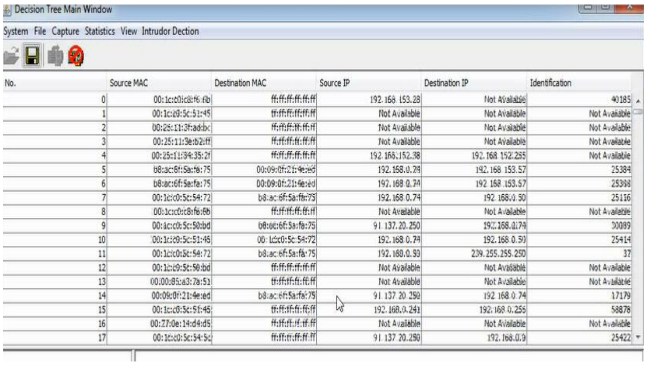
	SENSITIVITY	ACCURACY
ID3	96%	95%
J48	94.2%	97%

6. CONCLUSION

In order to protect web server, as a security tool, the intrusion detection system is indispensable. The GMM technique has been introduced to apply in the classification of rule set so as to improve the traditional classification technique, reduce the matching times and eventually improve the detection efficiency. In this paper we proposed a novel method based on Hierarchical Gaussian Mixture Model for intrusion detection mechanism. HGMM is an effective model for detecting computer attacks of unknown patterns. The Expectation-maximization algorithm are used to compute the parameters of a parametric mixture model distribution. If the threshold value is made too low, the IDS Engine suffers from a high false alarm rate. Here new scan detection techniques that have much lower false alarm rate and much higher coverage than existing techniques are used to reduce the overall false alarm rate. Some of the methods used are Filtering the unwanted packets and Setting medium level of threshold value.

Using Dimensionality Reduction for three dimensionalities such as for 41 attributes 14 attributes and 7 attributes the classification of attacks are made and by applying the evaluation criteria the corresponding Specificity, Accuracy, Sensitivity are evaluated to get the respective True Positive, false positive rate for both the algorithms .

SNORT RESULTS:



Decision Tree Main Window
System File Capture Statistics View Intrudor Dection

No.	Source MAC	Destination MAC	Source IP	Destination IP	Identification
0	00:1c:c0:c8:f6:6b	ff:ff:ff:ff:ff:ff	192.168.153.28	Not Available	40185
1	00:1c:c0:5c:51:45	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
2	00:25:11:3f:ad:bc	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
3	00:25:11:3e:b2:ff	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
4	00:25:11:34:35:2f	ff:ff:ff:ff:ff:ff	192.168.152.38	192.168.152.255	Not Available
5	b8:ac:6f:5a:fa:75	00:09:0f:21:4e:ed	192.168.0.74	192.168.153.57	25384
6	b8:ac:6f:5a:fa:75	00:09:0f:21:4e:ed	192.168.0.74	192.168.153.57	25368
7	00:1c:c0:5c:54:72	b8:ac:6f:5a:fa:75	192.168.0.74	192.168.0.50	25116
8	00:1c:c0:c8:f6:6b	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
9	00:1c:c0:5c:50:bd	b8:ac:6f:5a:fa:75	91.137.20.250	192.168.0.74	30089
10	00:1c:c0:5c:51:45	00:1c:c0:5c:54:72	192.168.0.74	192.168.0.50	25414
11	00:1c:c0:5c:54:72	b8:ac:6f:5a:fa:75	192.168.0.50	239.255.255.250	37
12	00:1c:c0:5c:50:bd	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
13	00:00:85:a3:7a:51	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
14	00:09:0f:21:4e:ed	b8:ac:6f:5a:fa:75	91.137.20.250	192.168.0.74	17179
15	00:1c:c0:5c:51:45	ff:ff:ff:ff:ff:ff	192.168.0.241	192.168.0.255	58878
16	00:27:0e:14:d4:d5	ff:ff:ff:ff:ff:ff	Not Available	Not Available	Not Available
17	00:1c:c0:5c:54:5c	ff:ff:ff:ff:ff:ff	91.137.20.250	192.168.0.9	25422

7. REFERENCES

[1] B. Woodward, R. S. H. Istepanian, and C. I. Richards, "Design of a telemedicine system using a mobile telephone", IEEE Trans. on Information Technology in Biomedicine, vol. 5, no. 1, pp. 13–15, March. 2001.

[2] Jinwook C., Sooyoung Y., Heekyong P., and Jonghoon C., "MobileMed: A PDA-based mobile clinical information system", IEEE Trans. on Information Technology in Biomedicine, vol. 10, no. 3, July 2006.

[3] Kyriacou E., S. Voskarides, C.S. Pattichis, R. Istepanian, M.S. Pattichis, C.N. Schizas, "Wireless Telemedicine Systems: A brief Overview", 4th International workshop on Enterprise Networking and Computing in Healthcare Industry (HEALTHCOM2002), Vol. 1, pp. 50-56, Nancy, France, June 2002.

[4] Lo B., Thiemjarus S., King R., and Yang G., "Body Sensor Network - A Wireless Sensor Platform for Pervasive Healthcare Monitoring", Adjunct Proceedings of the 3rd International conference on Pervasive Computing (PERVASIVE'05), May 2005.

[5] Milazzo Jr. A.S., Herlong J.R., Li J.S., Sanders S. P., Barrington M., and Bengur A.R., "Real-time transmission of pediatric echocardiograms using a single ISDN line", Computers in Biology and Medicine, vol. 32, pp. 379-388, September 2002.

[6] N. F. Timmons, W. G. Scanlon, "Analysis of the performance of IEEE 802.15.4 for medical sensor body area networking", IEEE Sensor and Ad Hoc Communications and Networks Conference (SECON), 2004.

[7] N. Smith-Guerin, L. Al Bassit, G. Poisson, C. Delgorge, P. Arbeille, and P. Vieyres, "Clinical validation of a mobile patient-expert tele-echography system using ISDN lines", in Proc. 4th Int. IEEE/EMBS Special Topic Conf. Inform. Technol. Applicat. Biomed., Birmingham, U.K., April 2003, pp. 23–26.

[8] Pertersen S., Peto V. and Rayner M., "Coronary heart disease statistics 2004", British Heart Foundation, June 2004

[9] R. S. H. Istepanian, E. Jovanov, Y. T. Zhang, "Guest editorial introduction to the special section on M-health: beyond seamless mobility and global wireless health-care connectivity", IEEE Trans. on Information Technology in Biomedicine, vol. 8, no. 4, December 2004.

[10] R. S. H. Istepanian, B. Woodward, and C. I. Richards, "Advances in telemedicine using mobile communications", in Proc. 23rd Annu. Int. IEEE/EMBS Conf., Istanbul, Turkey, 2001, pp. 3556–3558.

[11] Sinem Coleri Ergen, "Zigbee/IEEE 802.15.4 Summary", UC Berkeley, September 2004. http://www.cs.wisc.edu/~suman/courses/838/papers/zigbee.pdf

[12] V. Shnayder, B. Chen, "Sensor networks for medical care", Technical Report TR-08-05, Division of Engineering and Applied Science, Harvard University, 2005.http://www.eecs.harvard.edu/~brchen/papers/codebluetechrept05.pdf

[13] W. J. Tompkins, Ed., Biomedical Digital Signal Processing. London, U.K: Prentice-Hall, 1993.

[14] Yuechun Chu and Aura Ganz, "A mobile teletrauma system using 3G networks", IEEE Trans. on Information Technology in Biomedicine, vol. 8, no. 4, December 2004

[15] mHealth: The effectiveness of semantic healthcare knowledge frame work for Health Monitoring System Using Smart Phones : Onkar S Kemkar, Dr P B Dahikar, NSI-35, Belgaum 2011