Intrusion Detection in Big Data Era: A Multi-Objective Approach for Longer Model Lifespans

Dr C K Gomathy-Assistant Professor, Department of CSE, SCSVMV Deemed to be University, India

Mr.Sattu Sai Balaji, Mr.Peteti Mukesh, Mr.Samudrala Ajay Kumar, Mr.Settemoni Manohar

UG Scholars, Department of CSE, SCSVMV Deemed to be University, India

ABSTRACT:

Despite highly accurate intrusion detection schemes based on machine learning (ML) reported in the literature, changes in network traffic behaviour quickly yield low accuracy rates. An intrusion detection model update is not easily feasible due to the enormous amount of network traffic to be processed in near real-time for high-speed networks, in particular, under big data settings. In this paper, we propose a new scalable long-lasting intrusion detection architecture for the processing of network content and the building of a reliable ML-based intrusion detection model. Experiments performed through the analysis of five years of network traffic, about 20 TB of data, have shown that our approach extends the lifespan of our model by up to six weeks. That occurs because the average accuracy rate of our proposal lasted eight weeks after the training phase, and traditional ones reached only two weeks after the model building. Additionally, our proposal achieves up to 10 Gbps of detection throughput in a 20-core big data processing cluster.

KEYWORDS: Machine Learning, Intrusion Detection, Big Data, High-Speed Networks.

I.INTRODUCTION

In recent years, network devices have been significantly increasing their bandwidth capability. Therefore, the average broadband speeds are expected to double within only five years, growing from 39 Mbps in 2017 to 75.4 Mbps in 2022

Cyberattacks have also significantly increased their capabilities, e.g., in 2017, a Distributed-Denial-of-Service (DDoS) attack peaked at 600 Gbps – a year after it reached over 1.7 Tbps, which is a 183 percent throughput increase. Thus, when DDoS attacks are occurring, they might represent up to 25 percent of a country's total current Internet traffic. Hence, current and future deployment of Intrusion Detection System (IDS) mechanisms must be able to perform at such high-speed network bandwidths.

Traditionally, IDS techniques were built using signature based approaches, meaning that the cyberattacks are detected by matching a signature (e.g., well-known streaming of bits or a sequence of events). Therefore, only known attacks can be detected through such an approach. In addition, as new attacks are discovered over time, new signatures must be built demanding human intervention, and detection throughput is further decreased as more signatures must be evaluated [2]. Therefore, the detection can only be performed after the cyberattack occurrence.

II.INTRUSION DETECTION AND BIG DATA

Big data settings pose significant challenges to traditional intrusion detection mechanisms. This section further describes the typical ML-based intrusion detection schemes and how big data impacts them.

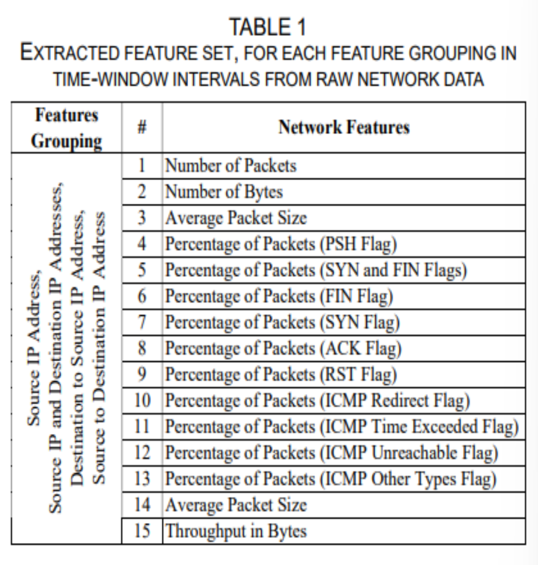


TABLE 1
EXTRACTED FEATURE SET, FOR EACH FEATURE GROUPING IN TIME-WINDOW INTERVALS FROM RAW NETWORK DATA

Features Grouping	#	Network Features
Source IP Address, Source IP and Destination IP Addresses, Destination to Source IP Address, Source to Destination IP Address	1	Number of Packets
	2	Number of Bytes
	3	Average Packet Size
	4	Percentage of Packets (PSH Flag)
	5	Percentage of Packets (SYN and FIN Flags)
	6	Percentage of Packets (FIN Flag)
	7	Percentage of Packets (SYN Flag)
	8	Percentage of Packets (ACK Flag)
	9	Percentage of Packets (RST Flag)
	10	Percentage of Packets (ICMP Redirect Flag)
	11	Percentage of Packets (ICMP Time Exceeded Flag)
	12	Percentage of Packets (ICMP Unreachable Flag)
	13	Percentage of Packets (ICMP Other Types Flag)
	14	Average Packet Size
	15	Throughput in Bytes

Extraction of intrusion behaviour from network data

Network-based intrusion detection systems (NIDSs) perform detection according to the intruder behaviour gathered from the network data content. For instance, network data content can be made of packets or network logs, such as NetFlow records, among others. In general, a huge amount of network packets (big data settings) arrives in a disorderly manner. In other words, the network packets must be pre-processed before being handled by a NIDS engine.

Machine learning for network-based intrusion detection

In general, intrusion detection through ML-based techniques is performed employing pattern recognition approaches [9], which have a goal of classifying a given input into a set of classes.

Network-based intrusion detection in big data settings

Big data scenarios are often characterized in 5 main aspects, namely 5Vs, which includes Volume, Velocity, Variety, Veracity, and Value [35]. For instance, consider a monitored high-speed network environment. In such a case, network data is generated at high velocity, which produces a vast amount of volume. The monitored network data may arrive in a variety of formats, which includes network packets, NetFlow records, or even application logs. Finally, the analysis of such data provides value, for instance, through the identification of an intrusion, if its veracity is assured. In such settings, traditional computing architectures are unable to cope with the processing demands [36]. Hence, big data environments require novel and distributed processing architectures, such as those provided on Hadoop ecosystem [5].

III.LIFESPAN OF TRADITIONAL MACHINE LEARNING DETECTION TECHNIQUES

Although the need for model updates for NIDS is a known requirement, the lifespan of current detection models remains unknown. This section evaluates the accuracy of degradation and the model lifespan of traditional ML detection techniques.

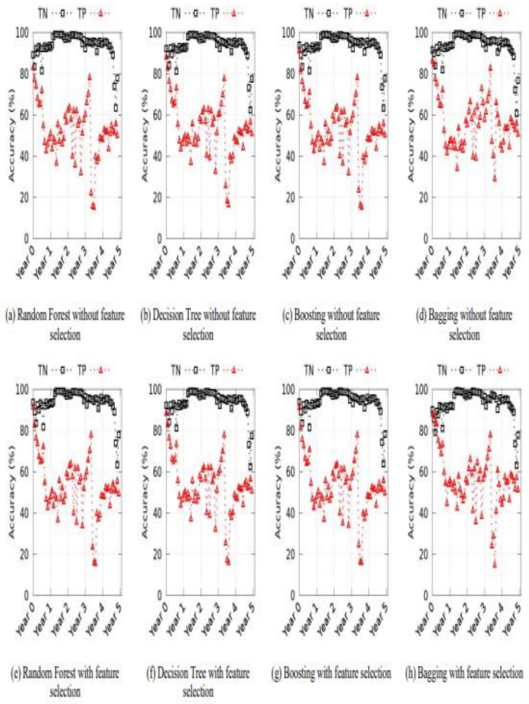
Data description

An important issue to be considered in intrusion modelling is to have a properly built training and testing dataset. A dataset used for such a purpose must be made of network data with real, valid, variable, publicly available, and correctly labelled events (network packets). However, in general, to provide such an enriched dataset, one must record real data, making data sharing unfeasible due to privacy concerns. Nonetheless, the evaluation of a model lifespan is even more difficult since data must be recorded for long periods, increasing the amount of data to be labelled and stored.

Accuracy behaviour over time

The first evaluation aims at assessing the ML model accuracy over time through the built dataset. Due to the imbalanced nature of the dataset (only ~2% of instances are samples of attack), a random under-sampling without replacement was performed in the training data. Hence, the data distribution used for training purposes is equally distributed between the classes.

FIGURE 1
ACCURACY BEHAVIOR OF SEVERAL ML ALGORITHMS WITH AND WITHOUT FEATURE SELECTION OVER A 5-YEAR RANGE.



(a) Random Forest without feature selection
(b) Decision Tree without feature selection
(c) Boosting without feature selection
(d) Bagging without feature selection
(e) Random Forest with feature selection
(f) Decision Tree with feature selection
(g) Boosting with feature selection
(h) Bagging with feature selection

FIGURE 2
RF ACCURACY TRADEOFF OVER TIME WITH AND WITHOUT FEATURE SELECTION

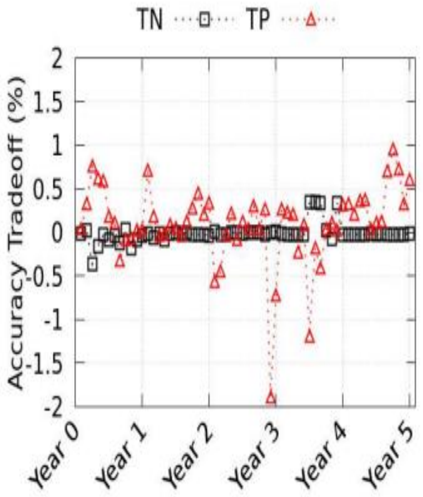
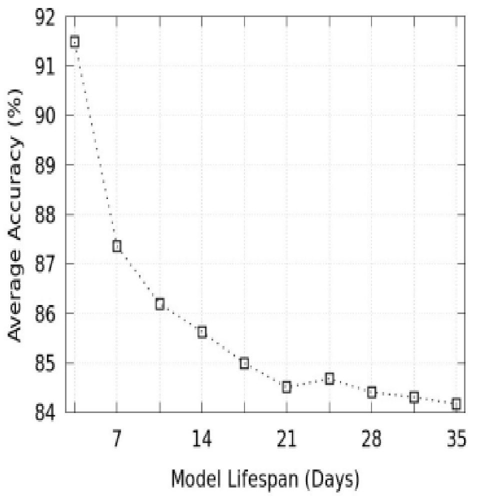


FIGURE 3
MODEL LIFESPAN (UPDATE PERIODICITY) AND AVERAGE ACCURACY FOR AN RF CLASSIFIER IN 2016



Discussion

Over the last few years, proposed ML-based intrusion detection schemes have disregarded the challenge of network traffic changes over time. This built dataset is a breakthrough toward the proper evaluation of ML-based intrusion detection schemes. To the best of our knowledge, it is the first dataset made of real network traffic, previously labelled, publicly available, and comprising many years of network traffic behaviour.

IV.LONG-LASTING INTRUSION DETECTION ARCHITECTURE FOR BIG DATA ENVIRONMENTS

In order to address the evolving behaviour of network traffic over time while also performing the intrusion in near real-time in big data environments, we propose a Long-Lasting Intrusion Detection Model. The proposed model is implemented in a twofold manner.

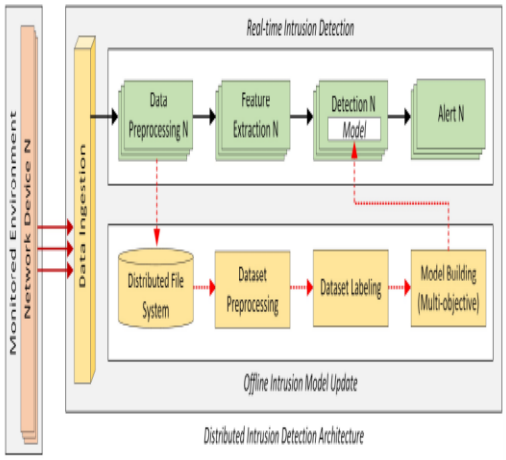
Architecture data flow

The proposed architecture acquires network data from monitored network devices (Figure 4, Network Device), such as routers, switches, servers, and other network hardware or devices exposed to network attacks.

Real-time Intrusion Detection

Despite the need for an ML model that properly classifies network events, the alarms must be generated on time, aiming to enable the proper handling of intrusion attempts while also helping reduce the damage when an attack is occurring. Therefore, ingested data must be classified as soon as possible by the Real-time Intrusion Detection module, which aims to detect network attacks in near real-time. The module is deployed as a stream processing big data architecture. The processing flow is performed as an acyclic graph flow. Each module is deployed in several processing worker nodes. Thus, the processing flow becomes scalable, providing a higher detection throughput.

FIGURE 4
LONG-LASTING INTRUSION DETECTION ARCHITECTURE FOR BIG DATA SETTINGS



Offline Intrusion Model Update

In general, the ML training phase demands a great amount of time to properly learn behaviour from the input event samples. In addition, in big data settings, the demanded time and required processing infrastructure for training significantly increase. The processing infrastructure for training involves ML algorithms, parameter optimization, feature selection, and model testing. Besides the execution of the IDS in near Realtime, the ML model must be periodically updated offline. Consequently, the Offline Intrusion Model Update module is deployed as a batch processing architecture. The processing flow is executed sequentially and divided among several processing worker nodes.

Building models with longer lifespans

In the state-of-art, current model building techniques do not take into account the model lifespan at the training phase. That happens mainly because proposals in the literature do not have a properly built training dataset that spans for long periods (see Section III). In contrast, our architecture evaluates the model lifespan and takes it into account during the feature selection task. However, to provide a high detection accuracy, during feature selection, the model lifespan is coupled with detection accuracy in a multi-objective feature selection process.

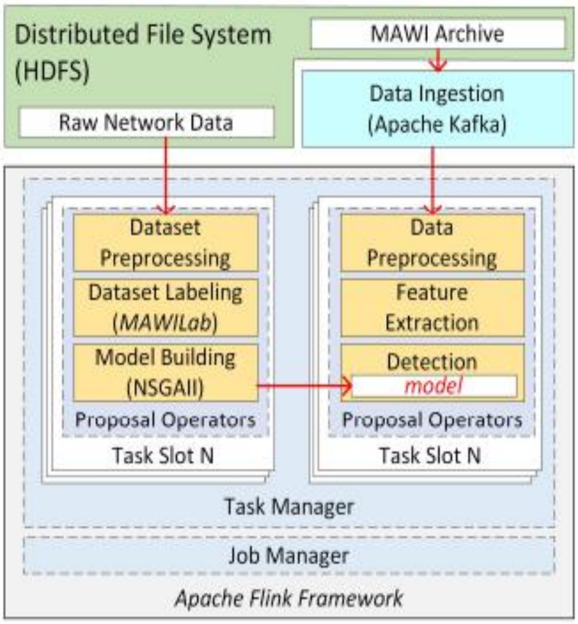
Discussion

Current approaches for network-based IDSs are not able to cope with the network traffic behaviour changes over time (Section III) and high-speed networks. In fact, in general, the model lifespan is not even evaluated. As a consequence, when ML-based techniques are deployed in production environments, the ML model quickly becomes outdated, and a new model must be built, which demands expert intervention and wastes time.

V.PROTOTYPE

A proposal prototype was implemented and deployed in a distributed environment, as shown in Figure 5. The prototype is implemented on top of Apache Flink [31] processing framework, version 1.8.1, due to its capability to operate in both batch and stream processing conditions.

FIGURE 5
PROPOSAL PROTOTYPE ARCHITECTURE



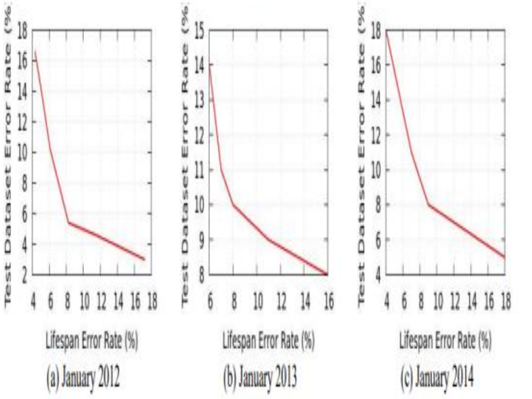
VI.EVALUATION

The evaluation of our proposal was performed in two steps. First, we evaluated the technique for multi-objective feature selection that aims for a higher model lifespan. Second, we evaluated the scalability.

Accuracy and Model Lifespan

The first evaluation comprises the accuracy and model lifespan improvement. To evaluate our proposed model, only the RF classifier was used, through the dataset introduced in Section III, as the other evaluated classifiers presented similar results. Similarly, the same set of parameters from Section III were used. Thus, the RF used 100 decision trees as its base learner.

FIGURE 6
MULTI-OBJECTIVE FEATURE SELECTION TRADEOFF BETWEEN MODEL LIFESPAN AND MODEL ACCURACY FOR THE THREE FIRST YEARS IN THE BUILT DATASET



(a) January 2012 (b) January 2013 (c) January 2014

FIGURE 7
MODEL LIFESPAN AND ACCURACY THROUGHOUT TIME

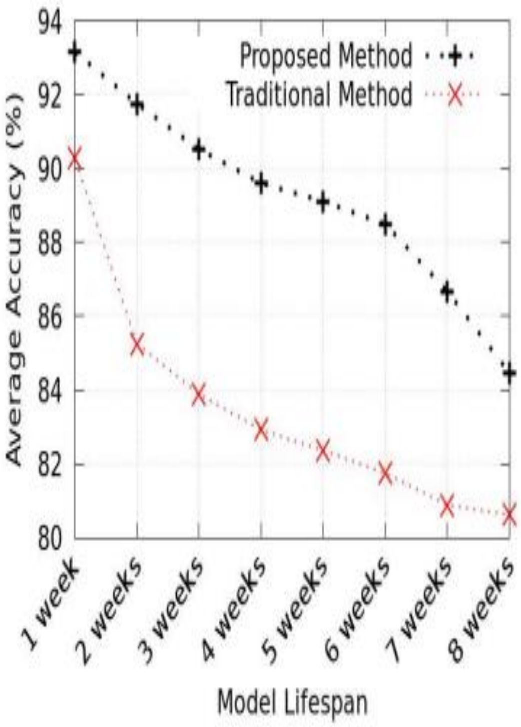


FIGURE 8
RF ACCURACY THROUGHOUT 5 YEARS WITH MONTHLY UPDATES AND MULTI-OBJECTIVE FEATURE SELECTION

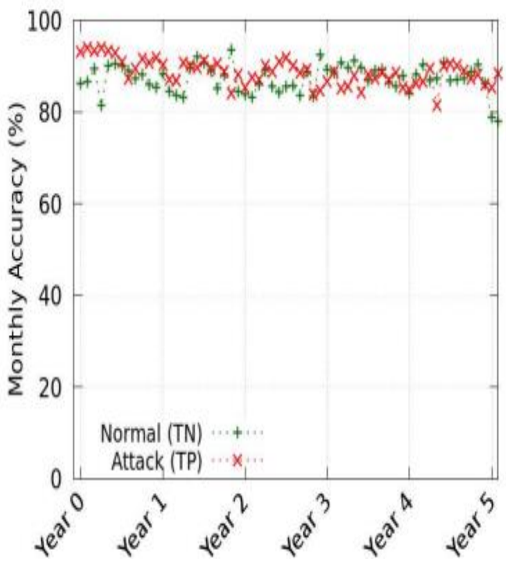
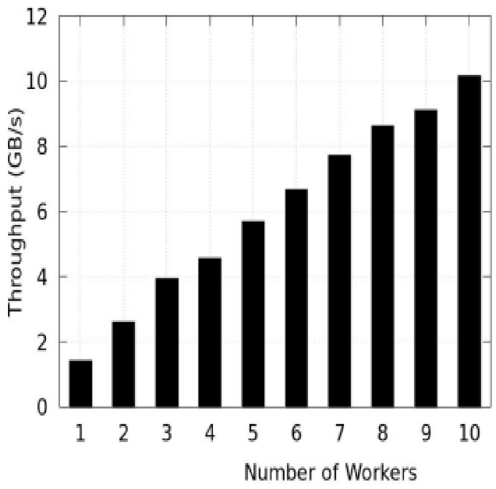


FIGURE 9
PROPOSAL SCALABILITY ACCORDING TO THE NUMBER OF USED WORKER NODES



Scalability:

To answer question (iv), the architecture throughput was evaluated during deployment using our 11-node cluster. The classification throughput was evaluated according to the Realtime Intrusion Detection module, considering that the model update task is typically performed offline. The architecture was deployed using 11 nodes, each with a 2-core CPU, 8 GB of memory, and an Ubuntu 18.04 OS. Out of the 11 nodes, ten were used as worker nodes (Figure 5, Task Manager) while the remaining node was responsible for the infrastructure management, acting as the master node (Figure 5, Job Manager). For each evaluation, the architecture was executed for 30 minutes, while its throughput was measured according to the network packet ingestion rate (Figure 5, Data Ingestion to Data Pre-processing)

VII.CONCLUSION AND FUTURE WORK

In recent years, the state-of-the-art of the model update task in ML-based IDSs has been neglected by the research community. In this work, we have tackled the problem of ML model's lifespan in big data environments.

IX.REFERENCES

[1] DR.C.K.Gomathy , V.Geetha , S.Madhumitha , S.Sangeetha , R.Vishnupriya Article: A Secure With Efficient Data Transaction In Cloud Service, Published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 4, March 2016, ISSN: 2278 – 1323.

[2] Dr.C.K.Gomathy,C K Hemalatha, Article: A Study On Employee Safety And Health Management International Research Journal Of Engineering And Technology (Irjet)- Volume: 08 Issue: 04 | Apr 2021

[3] Dr.C K Gomathy, Article: A Study on the Effect of Digital Literacy and information Management, IAETSD Journal For Advanced Research In Applied Sciences, Volume 7 Issue 3, P.No-51-57, ISSN NO: 2279-543X,Mar/2018

[4] Dr.C K Gomathy, Article: An Effective Innovation Technology In Enhancing Teaching And Learning Of Knowledge Using Ict Methods, International Journal Of

Contemporary Research In Computer Science And Technology (Ijcrcst) E-Issn: 2395-5325 Volume3, Issue 4,P.No-10-13, April ’2017

[5] Dr.C K Gomathy, Article: Supply chain-Impact of importance and Technology in Software Release Management, International Journal of Scientific Research in Computer Science Engineering and Information Technology (IJSRCSEIT) Volume 3 | Issue 6 | ISSN : 2456-3307, P.No:1-4, July-2018.

[6] C K Gomathy and V Geetha. Article: A Real Time Analysis of Service based using Mobile Phone Controlled Vehicle using DTMF for Accident Prevention. International Journal of Computer Applications 138(2):11-13, March 2016. Published by Foundation of Computer Science (FCS), NY, USA,ISSN No: 0975-8887

[7] C K Gomathy and V Geetha. Article: Evaluation on Ethernet based Passive Optical Network Service Enhancement through Splitting of Architecture. International Journal of Computer Applications 138(2):14-17, March 2016. Published by Foundation of Computer Science (FCS), NY, USA, ISSN No: 0975-8887

[8] C.K.Gomathy and Dr.S.Rajalakshmi.(2014), "A Software Design Pattern for Bank Service Oriented Architecture", International Journal of Advanced Research in Computer Engineering and Technology(IJARCET), Volume 3,Issue IV, April 2014,P.No:1302-1306, ,ISSN:2278-1323.

[9] C. K. Gomathy and S. Rajalakshmi, "A software quality metric performance of professional management in service oriented architecture," Second International Conference on Current Trends in Engineering and Technology - ICCTET 2014, 2014, pp. 41-47, doi: 10.1109/ICCTET.2014.6966260.

[10] Dr.C K Gomathy, V Geetha ,T N V Siddartha, M Sandeep , B Srinivasa Srujay Article: Web Service Composition In A Digitalized Health Care Environment For Effective Communications, Published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 4, April 2016, ISSN: 2278 – 1323.

[11] C.K.Gomathy.(2010),"Cloud Computing: Business Management for Effective Service Oriented Architecture" International Journal of Power Control Signal and Computation (IJPCSC), Volume 1, Issue IV, Oct - Dec 2010, P.No:22-27, ISSN: 0976-268X .

[12] Dr.C K Gomathy, Article: A Study on the recent Advancements in Online Surveying , International Journal of Emerging technologies and Innovative Research (JETIR) Volume 5 | Issue 11 | ISSN : 2349-5162, P.No:327-331, Nov-2018

[13] Dr.C.K.Gomathy,C K Hemalatha, Article: A Study On Employee Safety And Health Management International Research Journal Of Engineering And Technology (Irjet)- Volume: 08 Issue: 04 | Apr 2021

[14] Dr.C K Gomathy, V Geetha , T.Jayanthi, M.Bhargavi, P.Sai Haritha Article: A Medical Information Security Using Cryptosystem For Wireless Sensor Networks, International Journal Of Contemporary Research In Computer Science And Technology (Ijcrcst) E-Issn: 2395-5325 Volume3, Issue 4, P.No-1-5,April ’2017

[15] C.K.Gomathy and Dr.S.Rajalakshmi.(2014), "Service Oriented Architecture to improve Quality of Software System in Public Sector Organization with Improved Progress Ability", Proceedings of ERCICA-2014, organized by Nitte Meenakshi Institute of Technology, Bangalore. Archived in Elsevier Xplore

Digital Library, August 2014, ISBN:978-9-3510-7216-4.

[16] Parameshwari, R. & Gomathy, C K. (2015). A Novel Approach to Identify Sullied Terms in Service Level Agreement. International Journal of Computer Applications. 115. 16-20. 10.5120/20163-2253.

[17] C.K.Gomathy and Dr.S.Rajalakshmi.(2014),"A Software Quality Metric Performance of Professional Management in Service Oriented Architecture", Proceedings of ICCTET'14, organized by Akshaya College of Engineering, Coimbatore. Archived in IEEE Xplore Digital Library, July 2014,ISBN:978-1-4799-7986-8.

[18] C.K.Gomathy and Dr.S.Rajalakshmi.(2011), "Business Process Development In Service Oriented Architecture", International Journal of Research in Computer Application and Management (IJRCM) ,Volume 1,Issue IV, August 2011,P.No:50-53,ISSN : 2231-1009

AUTHORS PROFILE:

S.Sai Balaji, B.E. Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

S.Manohar, B.E. Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

S.Ajay Kumar, B.E. Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

P.Mukesh, B.E. Computer Science and Engineering, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.

Dr.C.K.Gomathy, Assistant Professor, Computer Science and Engineering at Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India.