Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, July, 2021: 8386-8393

Research Article

A Survey on Automatic Abnormalities Monitoring System for Log Files using Machine Learning

Shivaprakash Ranga[1]*, Dr. Nageswara Guptha M[2], Dr. Hema M S[3]

Abstract

Attacks on a system's security are growing increasingly common since cyber criminals make a living out of system flaw exploitation. There is revenue loss and serious business and government impact. Widely used techniques to overcome these hazards are Signature recognition and anomaly detection, though these techniques are a good way to secure a system, they are unable to detect real time or modern attacks. The objective of this research is to survey different literature that uses security analytics to distinguish between malicious and normal activity. It also aims at creating a model that applies machine learning (ML) techniques to various log files which are at the server side with its nature as heterogeneous and can identify such users who are not given the access to these files using security analytics. The work proposes to develop and evaluate making use of the numerous production log files available on the website data.gov.in released under National Data Sharing and Accessibility Policy (NDSAP). This study establishes a foundation on which future research in this field may be constructed.

Introduction

Security attacks for a system are becoming more widespread since cyber attackers make use of system vulnerabilities for financial gain [1, 2]. For example, Ransom ware cyber-attack that swept the globe, India's largest container port in Mumbai that hit in June 2017 [3], the Visvesvaraya Technological University (VTU) student who hacked into the varsity site and pointed out major security holes in Oct 2017 [4] are the recent security breaches. The time taken to detect a security breach is measured in number of days [5] to make the situation still worse. These attackers employs various tools available in the underlying market like Rootkits, Malware Infection Frameworks (MIF) and Zero day exploits that allow them to circumvent standard security measures. These attackers can also launch additional attacks to acquire personal information (PI) and sensitive data. A security breach is unavoidable, but the greatest protection is to detect and correct assaults as soon as possible. To decrease the danger of a security breach, security experts employ preventive and detection measures.

[1]Research Scholar, Sri Venkateshwara College of Engineering, Bengaluru, India
[2]Professor, Sri Venkateshwara College of Engineering, Bengaluru, India
[3]Professor, Aurora's Scientific Technological & Research Academy, India
[1]shivaprakashranga@gmail.com, [2]mnguptha@yahoo.com, [3]ghema_shri@yahoo.co.in

A security breach is defined as “Any action the system owner deems unauthorized” [6]. Methods used to prevent this are focused on increasing the difficulty of attack. These techniques include establishment of a very effective protection policy, making use of current protection updates, fending off default configurations, and organizing a powerful person protection schooling program [7]. All statistics protection regulations ought to adhere to the 3 ideas of the CIA triad. They can be classified as Confidentiality, Integrity, and Availability. The former, “confidentiality”, refers to a collection of regulations that restricts data access. Integrity refers to the confidence that data is reliable and correct. The latter, “availability”, refers to the ability of permitted users to access information systems.

There are two types of detection methods. The detection are based on

I. Signature
II. Anomaly

Existing Techniques which focus on providing security such as virus scanner, Intrusion Detection Systems (IDS), Firewall which uses signature-based approach. This method compares a payload to a database consisting previous harmful signatures using hash technology [8]. Signature-based detection approaches look for current attacks in network data, but they can't identify the attacks which are modified and existing. These type of types are hard to detect and are commonly termed as mimicry attack. So the strategies should offer a robust protection towards attacks. However, they're never enough to defend against professional attackers who use today's assault strategies and exploits. Anomaly detection identifies unusual occurrences, including those that have never been seen before. Anomaly detection calls for a version of ordinary machine behavior. False alarms can arise while normal activities seem to be irregular. Of the best six objectives for future maintenance, development and research on cyber security, according to the Cyber Research Alliance (CRA), the most important is the employing Big Data Analytics. This is a relatively advanced trend in the business, expected to gain traction fast. Finding suitable algorithms in order to discover unseen patterns in big quantities of information is simply a challenge that needs to be addressed. Other aspects such as information containing noise and incomplete information are extra elements to be taken into consideration. Lastly, the big scale of organization safety information poses the finest challenge to successful implementation of Security Analytics. It varies from previous methods in that it distinguishes between what is normal and what is unusual. In other words, rather than the payload content or signature, the activity or user behavior is the focal point [10].

LITERATURE REVIEW

Log files are used by most systems to record events [11]. Their format and structure differ significantly depending on the platform as well as the system. For instance, servers are responsible for creating the logs. One of which is Apache server. Log files are used by OS, firewalls, and IDS to keep track of events. They are also used by applications to record user activity. Any activities involved in the breach of security will force in in creation of one or more log files.

However, a single log record event cannot detect these cyber-attacks, however a sequence of log entries covering many minutes can. The quantity of data logged each minute, per system might be in the hundreds of thousands. Furthermore, these files are often shared across a network. The log data requires merging and storage in a single location to be processed and analyzed. A huge centralized data repository is required to integrate extremely diverse data from different sources. The complexity criteria set out by Big Data should be met by such a data repository.

Dos was detected by an offline approach, which could also detect the assaults such as brute force password assaults [6]. It makes use of Anomaly detection as well as signature recognition for which have a signature database for attacks as well as regular signatures.

The set of rules collected in preprocessed log facts is used to identify more than one iteration of comparable log messages. Using this the device operates to search the private signature. This happens at the same time when the log files are processed. If a rule cluster detects any unfamiliar event taken place, the next process is to automate the process in which everyday log database is matched with the signature. It can be excluded if this turns out to be a true event. In case the current assault has the signature in it, then an alert is invoked.

By monitoring HTTP requests, Razzaq et al [8] presented a technique for identifying web application assaults. The outcome of this was to analyse the network traffic pattern by implementing it as a web proxy. The headers of users are examined and the HTTP protocol is analyzed. Further it also examines the user payload. The web application threats are identified he requests of users which implements the SQL infection. It is implemented by a model called as Ontology. It also includes attacks to poison the DNS cache. Well before the rules are processed on the web server, they are analyzed on the traffic of HTTP using requests of the users.

Postmortem intrusion detection [7] is employed to understand there was a intrusion, which systems were used to intrude and what part of the information was stolen. Here it is presupposed that the system is attacked by the intruder, bypassing any existing security measures such as an Intrusion Detection System.

In their paper "Network firewall utilising artificial neural networks," Valentan and Maly [12] use the back propagation approach to train a MLP ANN. With these, the firewall information are deduced by monitoring the network traffic. All of the systems evaluated have one major flaw: It lacks in identifying the attacks all together. Only the individual attacks on the subsystems are detected. So, most of the threats are detected with combining information from numerous log files rather than simply relying on a single log source. The attack can't be detected if any attack does not result in the creation of an event. [11].

Furthermore, there would be addition of overhead for systems which are operated to analyse the attacks over the raw network [13-15]. This will prevent data from reaching its destination on time. Intrusion Detection Systems and Firewalls are security measures that help to strengthen the network's security. These technologies should be supplemented by the implementation of less intrusive detecting techniques [16-36].

Proposed Work.

To develop and evaluate using a number of production log data supplied by the website data.gov.in released under National Data Sharing and Accessibility Policy [2] (NDSAP) and to develop a model which creates a model that applies machine learning (ML) techniques to various log files which are at the server side with its nature as heterogeneous and can identify such users who are not given the access to these files using security analytics.

The suggested method consists of five primary phases (Figure 1), with each step's result acting as feedback to give input information in the subsequent steps.

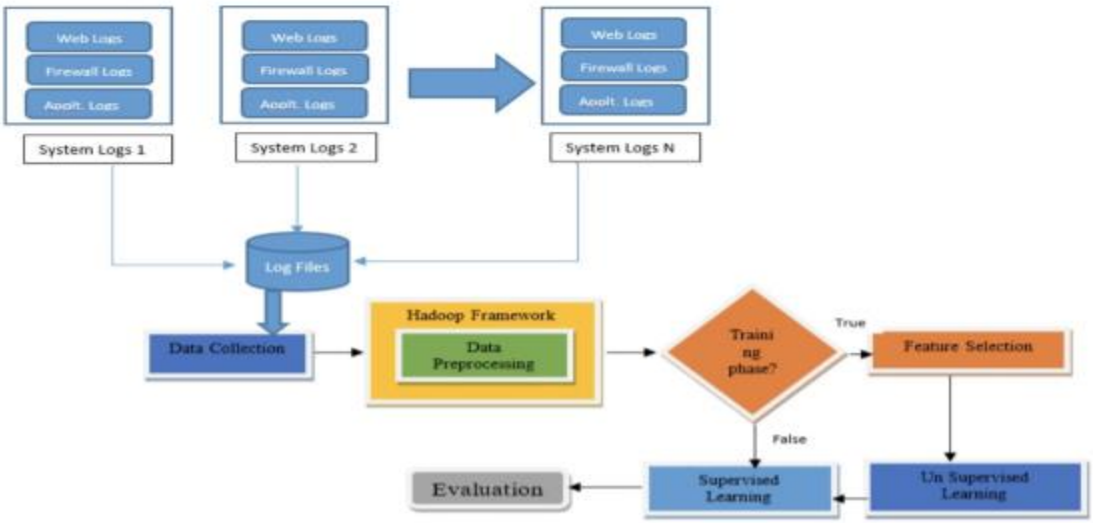
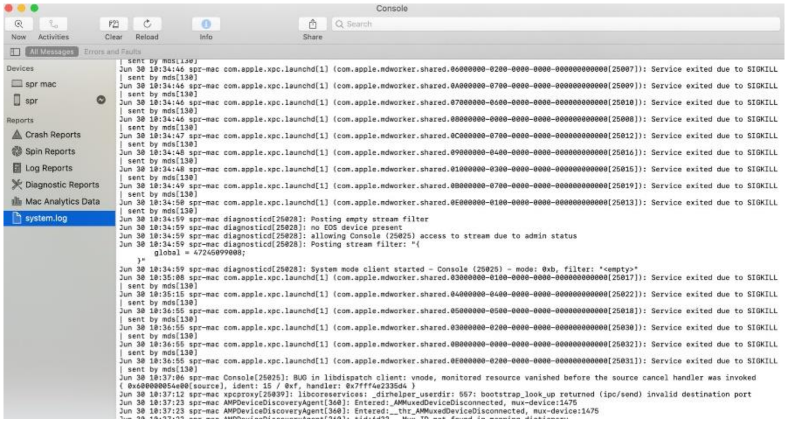


Figure 1: Process flow diagram

During data collection, the log data is collected from various systems in a connected network consisting of web logs, firewall logs, application logs and preprocessing is done to convert these inputs into usable type of format by ML techniques. In order to learn the patterns of how the user operates, unsupervised learning is employed. The clustering techniques assists this operation. The process of properly identifying the right features from the dataset which is pre-processed is known as feature selection. In supervised learning stage, the labelled dataset is useful in order to train the model. The model is trained once it delivers acceptable results and may be utilized in the production phase to detect anomalous user behavior.



```
| sent by mds[130]
Jun 30 10:34:46 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.06000000-0200-0000-0000-000000000000[25007]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:46 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.0A000000-0700-0000-0000-000000000000[25009]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:46 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.07000000-0600-0000-0000-000000000000[25010]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:46 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.08000000-0000-0000-0000-000000000000[25008]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:47 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.0C000000-0700-0000-0000-000000000000[25012]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:48 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.09000000-0400-0000-0000-000000000000[25016]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:48 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.01000000-0300-0000-0000-000000000000[25015]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:49 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.0B000000-0700-0000-0000-000000000000[25019]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:50 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.0E000000-0100-0000-0000-000000000000[25013]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:34:59 spr-mac diagnosticd[25028]: Posting empty stream filter
Jun 30 10:34:59 spr-mac diagnosticd[25028]: no EOS device present
Jun 30 10:34:59 spr-mac diagnosticd[25028]: allowing Console (25025) access to stream due to admin status
Jun 30 10:34:59 spr-mac diagnosticd[25028]: Posting stream filter: "{
        global = 47245099008;
    }"
Jun 30 10:34:59 spr-mac diagnosticd[25028]: System mode client started - Console (25025) - mode: 0xb, filter: "<empty>"
Jun 30 10:35:08 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.03000000-0100-0000-0000-000000000000[25017]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:35:15 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.04000000-0400-0000-0000-000000000000[25022]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:36:55 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.05000000-0500-0000-0000-000000000000[25018]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:36:55 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.03000000-0200-0000-0000-000000000000[25030]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:36:55 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.0B000000-0000-0000-0000-000000000000[25032]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:36:55 spr-mac com.apple.xpc.launchd[1] (com.apple.mdworker.shared.0E000000-0200-0000-0000-000000000000[25031]): Service exited due to SIGKILL
| sent by mds[130]
Jun 30 10:37:06 spr-mac Console[25025]: BUG in libdispatch client: vnode, monitored resource vanished before the source cancel handler was invoked
{ 0x600000054e00[source], ident: 15 / 0xf, handler: 0x7fff4e2335d4 }
Jun 30 10:37:12 spr-mac xpcproxy[25039]: libcoreservices: _dirhelper_userdir: 557: bootstrap_look_up returned (ipc/send) invalid destination port
Jun 30 10:37:23 spr-mac AMPDeviceDiscoveryAgent[360]: Entered:_AMMuxedDeviceDisconnected, mux-device:1475
Jun 30 10:37:23 spr-mac AMPDeviceDiscoveryAgent[360]: Entered:__thr_AMMuxedDeviceDisconnected, mux-device:1475
```

Figure 2: System Log File

The Figure 2 shows the system log file and various events captured during the login into the system. Similarly the various log files are collected and analyzed during the process. Since the data is collected from various systems and processed we can consider the data as big data and to process we can rely on the Apache Hadoop framework. If distributed processing is the need of the hour over the heap of data, then Hadoop software library called Apache is employed. It has a capability to develop as a whole of many computing systems from an individual server. This system has capabilities of computing which is independent of any other systems. The library is used to find the intrusion instead of using an expensive hardware. This library has the application layer fault handling capability. This allows the top of cluster of computer to get very good service. A suitable algorithm is applied to detect the suspicious user for example if the log file is unstructured, the algorithm like k-means, principal component analysis (PCA) is applied and for structured log files for classification between normal and abnormal user the algorithms like Random forest, K- Nearest Neighbor (KNN) is applied.

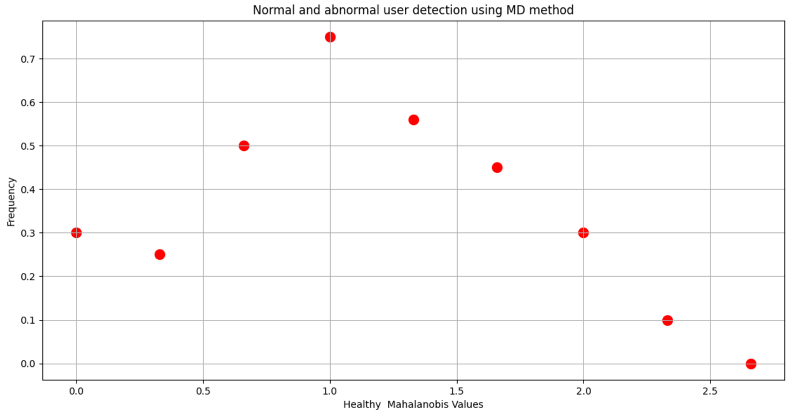


Figure 3: Normal and Abnormal user detection

A threshold value is defined to indicate the range of values that can decide it to be an anomaly. An anomaly is indicated by sigma value > 2.5 from the Figure 3.

Conclusion

The goal of this study is to review the literature on how security analytics can be used to distinguish between malicious and legitimate user activity, and to develop a model that uses security analytics to detect intrusion by applying ML techniques. These techniques are applied to multiple heterogeneous server-side log files the log files. The processing of these files are on the scale of terabyte every day, and for this a computing implementation such as Hadoop can process terabytes of log data.

The first step in these types of investigation is to detect a cyber-attack. Risk mitigation steps, such as blacklisting originating source IPs and freezing accounts, may be required by the security analyst. The IP address origin, data used by attacker to access the files, and the log files should be carefully analyzed in order to detect any accounts that are compromised. Future studies might include automated correlation and also the information to analyse the attacks of the log data.

Extra ML calculations and investigation implemented in this regard can lead to complexity to the functionality of the computing systems. This relates to the amount of data that has to be analyzed and also the speed at which the processing should take place.

References

1. Literature available at URL: www.data.gov.in/
2. Literature available at URL: https://dst.gov.in/
3. Literature available at URL: https://www.indiatoday.in/india/story/petya-ransomware-major-global-cyber-attack-wannacry-jawaharlal-nehru-port-trust-985106-2017-06-28
4. Literature available at URL: https://www.deccanherald.com/content/636138/vtu-not-amused-student-hacks.html (Web Link)
5. Phil Muncaster. "Hackers Spend Over 200 Days Inside Systems Before Discovery," Infosecurity Magazine. N.p., 24 Feb. 2015, https://www.infosecuritymagazine.com/news/hackersspend
6. J. Ng, D. Joshi, and S. M. Banik. "Applying Data Mining Techniques to Intrusion Detection," Information Technology: New Generations (ITNG) 2015 Proceedings of the 12th International Conference on Information Technology. April 13, 2015, LasVegas, NV, USA. pp. 800-801.
7. K. A. Garcia, et al. "Analyzing Log Files for Postmortem Intrusion Detection." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.6 (2012): pp 1690-704.
8. Razzaq, Abdul, et al. “Semantic Security Against Web Application Attacks, ” Vol 254. Elsevier Inc, 2014.
9. Kott, A., and C. Arnold. "The Promises and Challenges of Continuous Monitoring and Risk Scoring." IEEE Transactions on Security & Privacy 11.1 (2013): pp. 90-3.
10. Mahmood, T Mahmood, T., and U. Afzal. "Security Analytics: Big Data Analytics for Cybersecurity: A Review of Trends, Techniques and Tools". 2013 2nd National Conference on Information Assurance (NCIA). December, 11 2013, Rawalpindi, Pakistan. pp. 129-134.
11. Abad, C., Taylor, J., Sengul, C., Yurcik, W., Yuanyuan Zhou, & Rowe, K. “Log correlation for intrusion detection: A proof of concept.” ASCAC ’03 Proceedings of the 19th Annual Computer Security Applications Conference. December, 8, 2003, Las Vegas, NV, USA, pp. 255- 264.
12. Valent n, Kristi n, and Michal MALY. "Network Firewall using Artificial Neural Networks." Computing & Informatics 32.6 (2013): 1312-27.
13. S Kanagaraj, M S Hema, M. Nageswara Guptha (2019), “Machine Learning Techniques for Prediction of Parkinson’s Disease using Bigdata”, International Journal Recent Trend Technology and Engineering, Volume 8, No. 10, pp. 3788-3791.
14. Maheshprabhu, R. Hema, Srilatha Chepure, Nageswara Guptha M. Logistics Optimization in Supply Chain Management using Clustering Algorithms. Scalable Computing: Practice and Experience 2020, 21, 107–114.
15. M.S. Hema, R. Maheshprabhu, M. Nageswara Guptha, A. Sampathkumar, J. Amudhavel, Similarity learning of Parkinsonism elicited from genetics and syndrome for pharmacotherapy decision making, Materials Today: Proceedings, 2021
16. Ajay Sudhir Bale, Suhaas V. Reddy, Shivashankar A. Huddar, Electromechanical characterization of Nitinol based RF MEMS switch, Materials Today: Proceedings, Volume 27, Part 1, 2020, Pages 443-445, ISSN 2214-7853,

https://doi.org/10.1016/j.matpr.2019.11.263

17. Ajay Sudhir Bale, J. Aditya Khatokar, Shantanu Singh, G. Bharath, M.S. Kiran Mohan, Suhaas V. Reddy, T.Y. Satheesha, Shivashankar A. Huddar, Nanosciences fostering cross domain engineering applications, Materials Today: Proceedings, 2020, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.09.076.
18. J. Aditya Khatokar, N. Vinay, Ajay Sudhir Bale, M.A. Nayana, R. Harini, V. Suhaas Reddy, N. Soundarya, T.Y. Satheesha, A. Shivashankar Huddar, A study on improved methods in Micro-electromechanical systems technology, Materials Today: Proceedings, 2020, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.10.993.
19. S. A. Huddar, B. G. Sheeparamatti and A. S. Bale, "Study of pull-in voltage of a perforated SMA based MEMS Switch," 2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS), Vellore, India, 2017, pp. 1-4, doi: 10.1109/ICMDCS.2017.8211584.
20. Ajay Sudhir Bale et al 2020 IOP Conf. Ser.: Mater. Sci. Eng. 872 012008
21. Venkatesh M S, Manoj Patil, Ajay Sudhir Bale, Srujan Ingalgeri. Design of Remotely Monitorable Low Power Phototherapy Unit for Treatment of Neonatal Hyperbilirubinemia, National Conference at Bapuji Engineering College, Davangere, India
22. Aditya Khatokar J., Mounisha M., Nayana M.A., Ajay Sudhir Bale, Bhavana S. Battery Management System: A Survey. Journal of Industrial Safety Engineering. 2020; 7(1): 29–35p.
23. Kishan Das Menon H, Aditya Khatokar J, Ajay Sudhir Bale. Enhanced Railway Operations Using Automated Locomotive Simulator. Trends in Transport Engineering and Applications.2020; 7(1): 17–23p.
24. Ajay Sudhir Bale, Hosamani Ummar Farooq N, Shivashankar Huddar. Automated Diesel transfer system using PLC. Journal of Industrial Safety Engineering. 2019; 6(1): 8–14p.
25. Aditya Khatokar J, Nayana M A , Soundarya N, Meghana N, Bhavana S, Sunkireddy Umarani, Ajay Sudhir Bale. Electric Vehicles: Transition to Green Zone. Trends in Transport Engineering and Applications. 2020; 7(2): 12–17p.
26. Raksha K.P., Rajani Alagawadi, Nisha N., Deeksha R., Ajay Sudhir Bale. Advancement of Nanotechnology in Batteries. International Journal of Energetic Materials. 2020; 6(2): 18–24p.
27. Vinay N., Aditya Khatokar J., Ajay Sudhir Bale. Analysis on Synthesis of Quantum Dots with Their Applications on Photochemistry. International Journal of Photochemistry. 2020; 6(1): 1–11p
28. Ajay Sudhir Bale, Bharath G, Kiran Mohan M S, Shantanu Singh, Aditya Khatokar J. Thin-Films: Study of Medical, Display and Environmental Applications. International Journal of Energetic Materials. 2020; 6(1): 1–6p.
29. Aditya Khatokar J., Nayana M.A., Ajay Sudhir Bale, Meghana N., Sunkireddy Umarani. A Survey on High Frequency Radios and their Applications. Journal of Industrial Safety Engineering. 2020; 7(1):7–12p.
30. Harish Koujalgi, Ajay Sudhir Bale. Biometric Based Automatic Ticket Vending Machine for Indian Railways. International Research Journal of Engineering and Technology (IRJET). Volume: 04 Issue: 07 July -2017. e-ISSN: 2395-0056, p-ISSN: 2395-0072.
31. Ajay Sudhir Bale, Harish Koujalgi. Quality Factor analysis for Nitinol based RF MEMS Resonator. International Research Journal of Engineering and Technology (IRJET). Volume: 04 Issue: 07 July -2017. e-ISSN: 2395-0056, p-ISSN: 2395-0072.

32. Aditya Khatokar J, Nayana M A, Kishan Das Menon H, Janardhan V, Ajay Sudhir Bale. A Study on Various Approaches in Remote Sensing. Journal of Telecommunication, Switching Systems and Networks. 2020; 7(2): 32–37p.
33. Ajay Sudhir Bale, J. Aditya Khatokar, M.S. Kiran Mohan, G. Bharath, Shantanu Singh, J. Roshini, Suhaas V. Reddy, Shivashankar A. Huddar, N. Vinay, Nanotechnology as a tool for treating cancerous tumors, Materials Today: Proceedings,2021,ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.12.1175.
34. S. S. Kumar, A. Sudhir Bale, P. M. Matapati and V. N, "Conceptual Study of Artificial Intelligence in Smart Cities with Industry 4.0," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 575-577, doi: 10.1109/ICACITE51222.2021.9404607.
35. A. S. Bale, S. Saravana Kumar, P. Rao and A. K. J., "A Recent Trend in DC Microgrid," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering(ICACITE), 2021, pp. 543-546, doi: 10.1109/ICACITE51222.2021.9404668.
36. Ajay Sudhir Bale, Subhashish Tiwari, K. Lova Raju, Pravesh P., Kishore P., Vinayak N. (2021). Environmental Surveillance Monitoring System In Industries Using Industrial Internet Of Things. Design Engineering, 1783- 1790. Retrieved from http://www.thedesignengineering.com/index.php/DE/article/view/1884