Nile Journal of Communication & Computer Science

Volume 7, June 2024

Journal Webpage: https://njccs.journals.ekb.eg

Print ISSN: 2805-2366
Online ISSN: 2805-2374

A Framework for Confidential Document Leakage Detection and Prevention

Hesham. A. Sakr, Magda I. El-Afifi

Assistant Professor, ECE department, Nile Higher Institute of Engineering and Technology, Mansoura, Egypt
Artificial Intelligence Lab

Abstract

Nowadays, With the spread of information crimes, the protection of confidential and sensitive data and documents from leakage and publication has become a major concern for government and private agencies, in addition, with the increase of this phenomenon in conjunction with the spread of new social media, which gave way to its perpetration and harm to organizations by leaking information by their employees. In this research, we have to create a framework to secure and protect these confidential and sensitive documents from two sides. The first aspect is prevention and lies in avoiding and limiting this unacceptable behavior by taking appropriate special measures to protect documents, and on the other side, emergency responses and the discovery of data leakage as soon as possible, addressing it, and detecting the cause of this, as the digital forensic investigation provides the tools and techniques that expose and punish the author of this disgraceful act.

Keywords: Cybersecurity, information crime, leaking and publishing documents Data loss prevention (DLP), loss and data breach.

1. Introduction

Data security is a primary concern of all applications in distributed environments. Protecting sensitive and confidential data from being shared with others is deemed as an ultimate goal of any business. Many security practices are followed to maintain the confidentiality of data, which ensures that data is retained in accordance with security policies and rules. An insider assault depicts the harm that can happen to the interests of an association by an unsuspicious or trusted individual with real admittance to its organization and framework assets. Such an assault can happen through a coincidental security break by an approved client, arranged security break by an approved client, or by an untouchable through a compromised framework. An arranged insider assault can bring about the exfiltration or annihilation of delicate information or can think twice about correspondences organization and different organization servers with more assets. In the present generally associated network conditions, an effective insider assault could result in serious causing harm to a venture's interests [1-3]. Nowadays, many of the electronic communications

heavily used within any organization for many purposes, such as local mail, instant messaging, web mail, data file transferring, and also organization website, continue to go largely to different destinations with no limitations, monitoring, or control on its movements from the organization. As a result of this issue, there is a high risk that the organization's confidential information will get into the wrong hands. Clearly, from this key point, the organization's sensitive data should be extremely properly protected, or else it will face catastrophic outcomes such as: business loss, damaged reputation, poor publicity, loss of strategic customers, and loss of competition with other organizations.[7-4]

As a result, any organization using a similar electronic document system must keep a close eye on sensitive data that has passed through this system or application in order to maintain reputation and business continuity, as well as ensure regulatory and legal compliance, while remaining distinct from others. The Data Leakage Prevention (DLP) solution, which basically protects sensitive data of an organization from being read by inappropriate individuals, whether from outside or even inside the organization, has recently risen to the top. This essentially means that specific data can only be read by a limited number of authorised individuals or groups. Government departments that grant access to pooled data to designated employees are frequently defenceless against insider attacks. Private places, for example, medicine organizations that retain essential and extremely sensitive private data, and banking organizations that deal with the evolution of money related exchanges, are also somewhat vulnerable to insider assaults. In such cases, the consequences of an insider attack may become dire, resulting in a lack of money, public scepticism, and legal ramifications. Furthermore, with greater openness and accessibility of sensitive data, insider assaults are likely to increase, as demonstrated in (Fig.1).

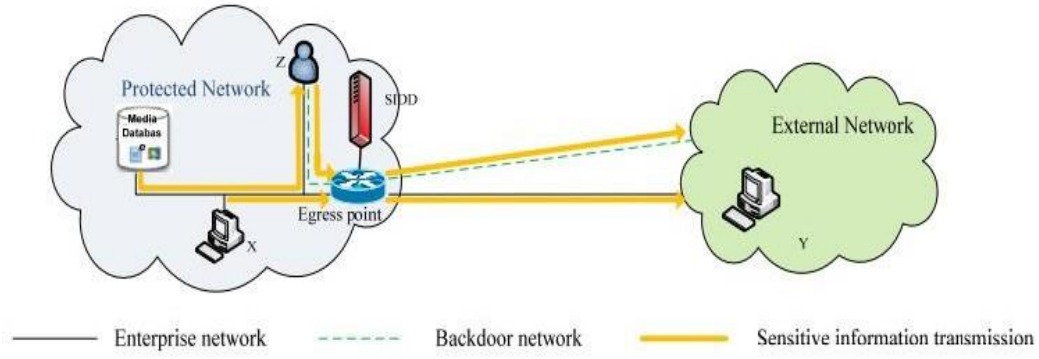

Figure 1 A motivating example for sensitive data exfiltration and detection

2. Related work

Yali Liu et al proposed a multilayer framework known as the Sensitive Information Dissemination Detection (SIDD) device, which is a high-speed obvious network bridge positioned at the included community's threshold. SIDD is comprised of three major components: 1) network-level application identification, 2) content material signature development and detection, and 3) covert communication detection. Furthermore, we present a prototype implementation of the key components, demonstrating how such a device could be deployed. Their strategy is entirely dependent on the use of statistics and sign processing procedures on site visitors waft to generate signatures and/or extract capabilities for classification reasons. Their proposed framework aims to deal with techniques to identify, dissuade, and save you intended and unintentional dissemination of sensitive content outside the organization using the organization's device and network resources via a trusted insider [8-11].

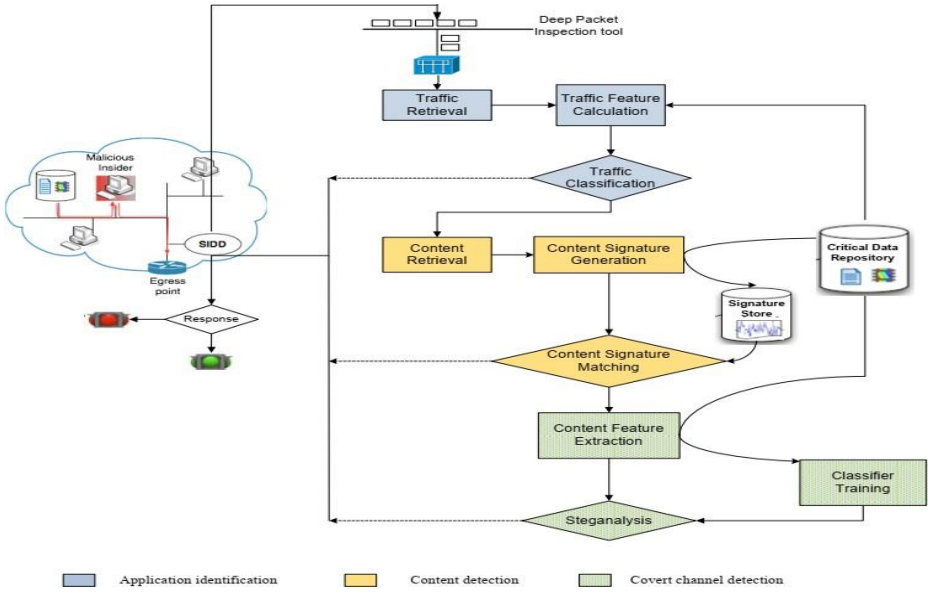

Figure 2 Sensitive Information Dissemination Detection (SIDD) algorithm.

Sensitive information in organizations might include company internal regulations, financial data, individual credit card statistics, and company-related records; these types of sensitive information can be released by a hostile person. Information leakage can cause major problems for a variety of organizations. A few Data Leakage Detection (DLD) versions used 'fake items' that were stored in the server database. The phoney things aid in luring and identifying the person who leaked the information. Everyone has the opportunity to leak the record, which is known as the guilt opportunity. Many detection models for record leakage awareness at the phoney goods as well as the database to find the leaker [11-14].

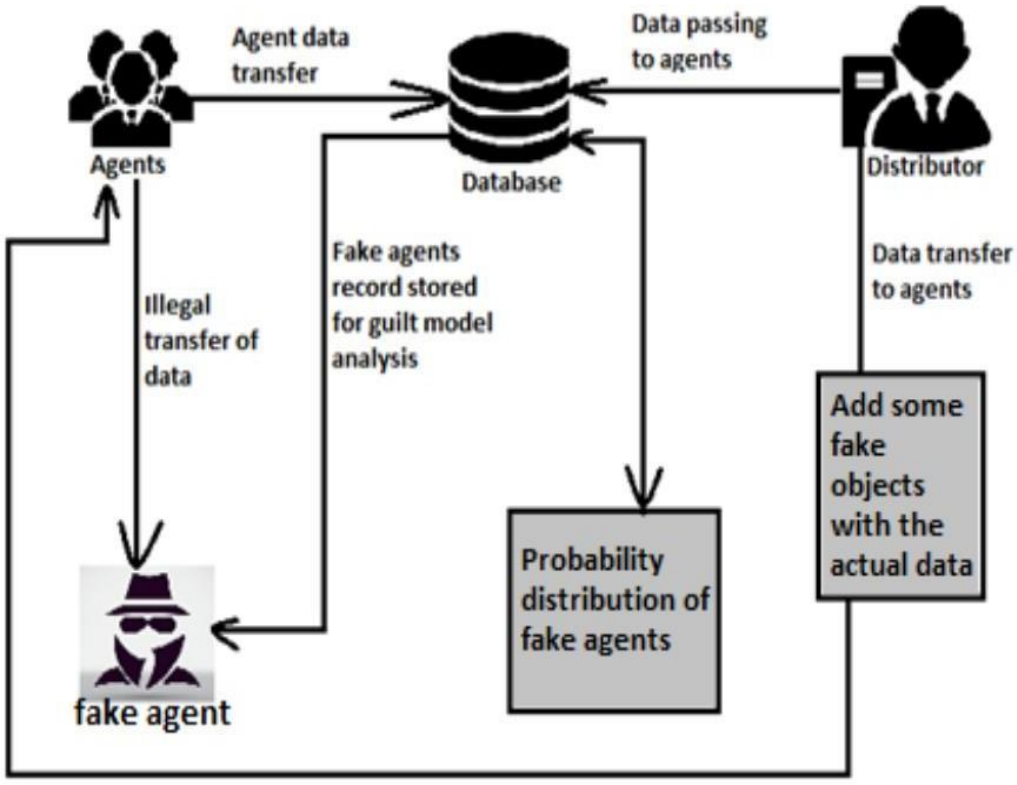

Figure 3 Process of a data leakage detection algorithm.

3. Contribution

Data leakage causes huge money and non-financial losses to those parties and is dangerous by which information is a crucial plus for these organizations as composed confidential data will seem in numerous leak channels and challenge the massive issue in protective info is the way to maintain the confidentiality of sensitive information. Given the implications of this disgraceful act and therefore the severe consequences and nice damages to government and personal agencies and their people. As a results, of revealing their secrets that may have many adverse effects on them and the reflection of the community's trust in them. As an organization's staff ought to access such data so as to hold out their daily work. Data leak detection is an essential and difficult task, whether or not caused by malicious intent or an accidental mistake, data loss may result in important damage to the organization. In this project, we explore an effective framework to protect these confidential and sensitive documents from two aspects: the first aspect is the prevention, which relies on avoiding and limiting unacceptable document leakage behavior by taking appropriate special measures and precautions to protect those documents, and on the other hand, considering the contingency responses and the discovery of data leakage as soon as possible, to promptly detected and addressed; and the second aspect, is to conduct a systematic review of previous relevant studies, taking into account a decade of current research efforts and highlighting issues and research gaps in these studies [15].

4. Techniques

- **Need for Digital Forensic**

due to the reality that the majority data loss is from inside personnel. The use of scientific testing or techniques in criminal investigations is known as forensics. The process of recovering and preserving materials found on digital devices is known as digital forensics. Data is frequently encrypted, erased, or buried, necessitating the use of digital forensics. Digital forensics is divided into five main areas based on where data is stored or how data is transferred. Digital forensics tools are hardware and software solutions that aid in the recovery and preservation of digital evidence. Digital forensics technologies can be used by law enforcement to collect and preserve digital evidence as well as to support or disprove hypotheses in court [16].

- **Digital Forensic Tools**

Digital forensics software and hardware are used interchangeably by law enforcement. Because computer and mobile device forensics are more popular, most technologies accessible to law enforcement, whether open source or commercial, focus on these two areas. Computers used in digital forensics have high performance requirements, needing larger capacity hard drives, quicker central processing units (CPUs), more memory, and so on.

- **Hardware**

Hardware tools are primarily intended for storage device investigations, with the goal of preserving the integrity of evidence by leaving suspect devices unchanged. A forensic disc controller, sometimes known as a hardware write blocker, is a read-only device that allows the user to access data from a questionable device without the risk of changing or wiping it. A disc write-protector, on the other hand, prevents the content of a storage device from being modified or wiped. A hard-drive duplicator is a device that replicates all files on a questionable hard disc to a clean drive; it can also duplicate data in flash drives or secure digital (SD) cards. A password recovery device attempts to crack password-protected storage devices using techniques such as brute-force or dictionary assaults [17]

- **Software**

The majority of forensic software tools are versatile and may do multiple tasks in a single application. Some apps are open source, allowing expert programmers to alter the code to match their specific demands while saving law enforcement money. Some can process numerous devices at the same time or manage various operating systems (for example, Windows and Linux). The capabilities of these programmes can be classified according to the disciplines of digital forensics used. Computer forensics software supplements law enforcement's hardware tools. While hardware solutions like write-blockers are primarily concerned with preserving evidence in a target device, software applications can capture and analyse digital evidence collected from the suspect device. Suspects frequently hide or erase their data or partition their computer hard drives, making evidence harder to find; nevertheless, forensic software packages can assist investigators in recovering this evidence. The Windows Registry keeps track of when, where, and how a file is created, renamed, viewed, moved, or destroyed, and some apps may

gather and analyse these traces. In short, digital forensics software can retrieve and investigate specific user activity [18-20].

5. Methodology

User ID and password which will assigned to the agent for logining to the system. The agent need to send records for requesting the distributor and distributor will check the request and send records to the agent via way of means of including faked item in the records allocation module. In agent guilt module we will test sending alert message to the distributor while the agent has percentage any private records.

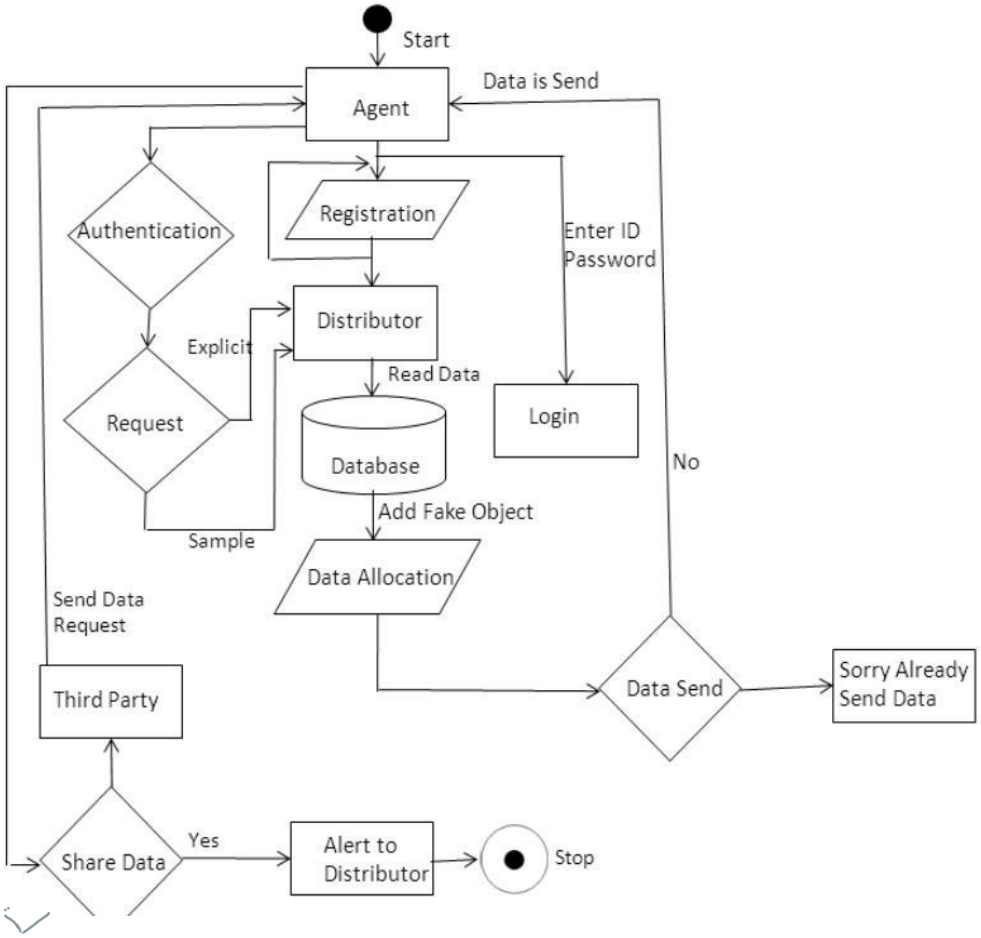

Figure 4 Data Loss Detection and prevention proposed (DLDP) algorithm.

The instructions of DLD and DLP techniques which has unique types of strategies to prevent statistics leakage and maintain the ownership and detection the use of behavioral assessment etc. The most well-known techniques beneath Neath DLP is the use of cryptographic and watermarking techniques, this can avoid unauthorized users access to the data. In the detection process, the

activity statistics and behavioral assessment with text mining have several likely developments as shown in figure 4,5 [21-25].

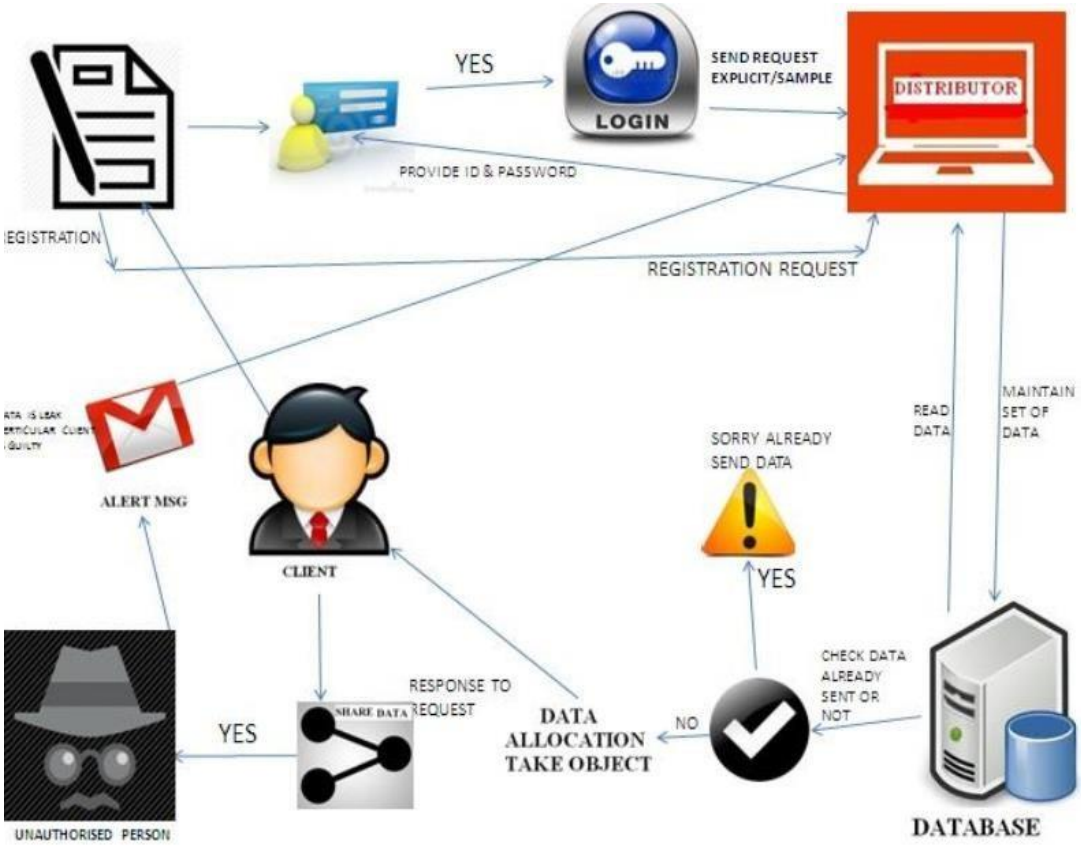

Figure 5. The Proposed Architecture Diagram.

6. Results

We applied a questionnaire which was conducted in this project to study a field case for employees working in the field of confidential documents in several sectors, including two organizations to which researchers belong to the project and choose that to find out the reasons for leaking confidential documents in different ages, specializations, countries and cultures, in order to find out the causes of this problem and what are the ways to overcome it In order to prevent leakage of confidential documents.

The questionnaire is divided into three sections as follows:

First section is interested with all general information about the employee such as :

1. Range of Age.
2. Years of Experience
3. Qualification
4. Social Status
5. Nationality.
6. One of the most widely used programs in point of view in illegally publishing confidential documents.
7. Have you been participating or attending short courses specialized in the area of the preservation of classified documents.
8. Are you aware and conscious of the sanctions' regime for the dissemination and disclosure of confidential documents and information issued by Royal Decree No. M/35 of 8 Jumada I 1432 AH?
9. Have you been made aware that some well-meaning and unintentional practices in using the electronic devices are considered as an information crime ?
10. Do you usually use easy to guess passwords like date of birth, phone number, or national ID ...etc?

The second section is interested with all general reasons about why secret documents were leaked and published according to the employee opinion such as:

1. The staff member's negligence in keeping confidential documents.
2. Exchange speeches among employees via modern social media.
3. existence of classified documents is unsafe, making it possible to leak them.
4. Leaving computers without login or screen lock to secure them.
5. Weak electronic systems make confidential documents easy to access and leak.
6. Unauthorized employees' access to and access to confidential documents.
7. hackers and hackers access the Intranet and leak data, most notably so-called social engineering, by luring employees and tricking them by external attackers to extract classified data.
8. Filming secret documents and publishing them to others in order to brag and brag.
9. The employee accessed suspicious websites and received suspicious messages or sent confidential documents from their personal mail.
10. Staff members have less awareness of how to handle confidential documents, of their own regulations and of the risks of document proliferation.
11. Their diversion in order to retaliate against an official and to offend the side for physical, functional, or dark reasons or for the detection of corruption.

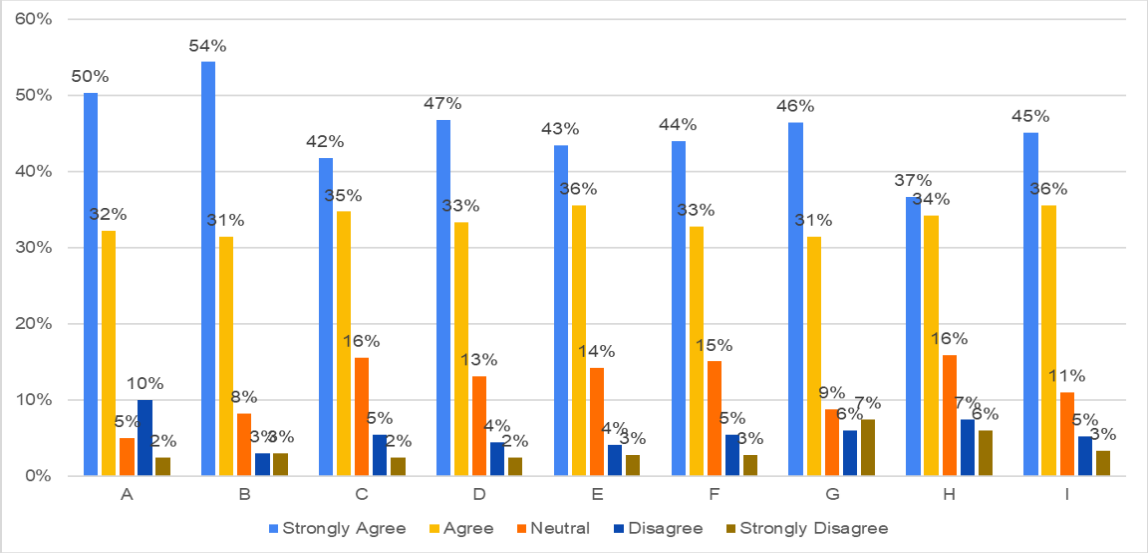

Figure 6. Reasons about why secret documents were leaked

Third section is divided into two parts the first one interested with all general reasons cause the publication and leaking of classified communications according to the employee opinion such as:

1. Abuse by hearing sectors and organizations.
2. Loss of trust in government or private systems.
3. Extortion of responsible personalities.
4. Subjecting them to forgery, misrepresentation and tampering with them to the excitement of society.
5. Leaking false classified documents aimed at offending responsible view.
6. It poses a threat to national security and damages the military, political, diplomatic, economic or social status.
7. The publication and diversion of confidential documents are either forged or tampered with and their content with a view to disrupting the work of vital sectors.

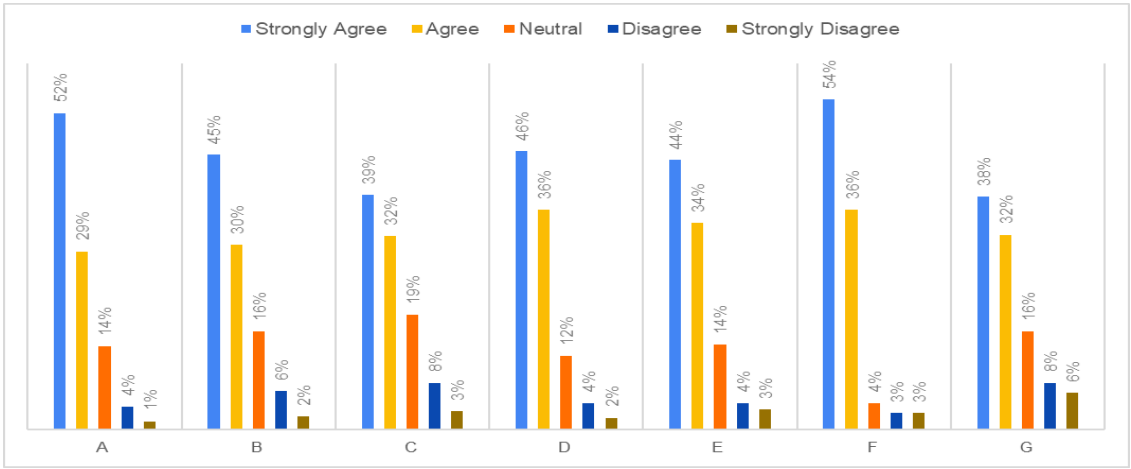

Figure 7 Causes the publication and leaking of classified communications

The second part interested with How to overcome the risk of publishing and leaking classified letters according to the employee opinion such as:

1. To educate and urge staff about the seriousness of this and to make them aware of the sale of security policies in the organizations.
2. A document of confidentiality and non-disclosure of information and the signature and readings of documents.
3. Activating and strengthening the strict sanctions regime against those found to have diverted any confidential letters.
4. Securing confidential documents by activating protection systems on all devices, networks and e-mails.
5. Classify confidential documents according to their importance and confidentiality in order to take appropriate action for each classification.
6. Use of confidential documents and correspondence by encryption devices to prevent hackers from accessing them.
7. Regulate the powers of access of employees to confidential data to be the powers according to the need.
8. The establishment of security controls in the workplace and the presence of confidential communications, such as access by staff and denial of Internet and other access.
9. To monitor and monitor the content of social media sites and the speed of its therapist's leak immediately by identifying to him and ways to deal with their diversion.
10. follow-up by government or private institutions to secure confidential documents and conduct training courses for sector employees to raise their awareness of the dangers of publishing and leaking documents.

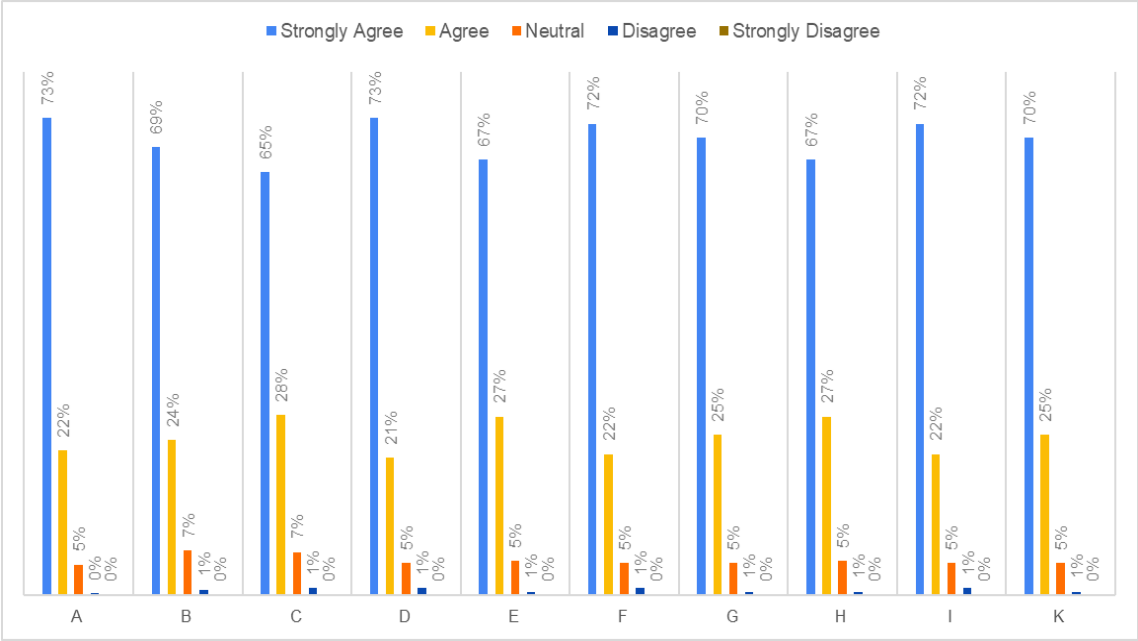

Figure 8 Risks about publishing and leaking

7. Conclusion

Based on human factors and location, there are two major reasons of data loss in an organization:

1. External Cause: Remote attackers are one of the most common external causes of data loss. Attackers are strangers who cannot get entry to organizations but can compromise them. Hacking the system with illicit software, implanted code, or social engineering attacks allows an organization to gain access to sensitive data and cause data loss.

2.Internal cause: Internal threats are one of the most common internal sources of data loss. from the inside out Threats are authorised personnel who can willfully abuse their powers and conduct maliciously and send critical data outside the organization's network [25-36].

8. Reference

1. M. R. Randazzo, M. Keeney, E. Kowalski, D. Cappelli, and A. Moore, “Insider Threat Study: Illicit Cyber Activity in the Banking and Finance Sector”, CERT and the National Threat Assessment Center, Aug. 2004.
2. E. D. Shaw, K. G. Ruby, and J. M. Post, “The insider threat to information systems: The psychology of the dangerous insider”, Security Awareness Bulletin, vol. 2-98, pp. 27-46, Sept. 1998.
3. L. Spitzner, “Honeypots: catching the insider threat”, Proceedings of 19th Annual Computer Security Applications Conference, pp. 170-179, Dec. 2003.
4. Liu, Yali, et al. "SIDD: A framework for detecting sensitive data exfiltration by an insider attack." 2009 42nd Hawaii international conference on system sciences. IEEE, 2009.
5. H.A.Sakr, and M.A.Mohamed, “Performance Evaluation Using Smart: HARQ Versus HARQ Mechanisms Beyond 5G Networks”, Wireless. Pers. Communication (Springer), pp.1-26, ISSN:1572-834X, June 2019.
6. Abeer Twakol Khalil, A. I. Abdel-Fatah and Hesham Ali sakr, “Rapidly IPv6 multimedia management schemes based LTE-A wireless networks", International Journal of Electrical and Computer Engineering (IJECE),vol. 9, no.pp. 3077-3089,2018.
7. H. A. Sakr, A. I. Abdel-Fatah, A. T. Khalil, “Performance Evaluation of Power Efficient Mechanisms on Multimedia over LTE-A Networks”, International Journal on Advanced Science, Engineering and Information Technology, (IJASEIT), vol. 9, no. 4, pp.1096-1109, 2019 .
8. H.A. Sakr and M.A. Mohamed, ‘Handover Management Optimization over LTE -A Network using S1 and X2 handover', Proc. of The Seventh International Conference on Advances in Computing, Electronics and Communication – ACEC 2018, ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-11, pp. 58–64, 2018.
9. Verma, Rajat, et al. "A Survey on Data Leakage Detection and Prevention." Proceedings of the International Conference on Innovative Computing & Communications (ICICC). 2020.
10. Elmrabit, Nebrase, Shuang-Hua Yang, and Lili Yang. "Insider threats in information security categories and approaches." 2015 21st International Conference on Automation and Computing (ICAC). IEEE, 2015.
11. Jadhav, Prasad, and P. M. Chawan. "Data Leak Prevention system: A Survey." Virus 6.10 (2019): 197-199.
12. Wadile, Kaveshwari R. "A Literature Review on Deta Leakage Detection." (2019).
13. M. Abdel-Azim, M., Awad, M. M., & Sakr, H. A.,” VoIP versus VoMPLS Performance Evaluation”, International Journal of Computer Science Issues (IJCSI), 11(1), 194, 2014.
14. M. Abdel-Azim, M., Awad, M. M., & Sakr, H. A.,” RSVP Based MPLS versus IP Performance Evaluation”, Mediterranean Journal of Computers and Networks (MEDJCN), 10(2), 2014.
15. Sakr, H. A., Ibrahim, H. M., & Khalil, A. T. (2022). Impact of Smart Power Efficient Modes on Multimedia Streaming Data Beyond 5G Networks. Wireless Personal Communications, 1-37.
16. Soomro, A. M., Naeem, A. B., Senapati, B., Bashir, K., Pradhan, S., Maaliw, R. R., & Sakr, H. A. (2023, January). Constructor Development: Predicting Object Communication Errors. In 2023 IEEE International Conference on Emerging Trends in Engineering, Sciences and Technology (ICES&T) (pp. 1-7). IEEE.
17. Soomro, A. M., Naeem, A. B., Senapati, B., Bashir, K., Pradhan, S., Ghafoor, M. I., & Sakr, H. A. (2023, January). In MANET: An Improved Hybrid Routing Approach for Disaster Management. In 2023 IEEE International Conference on Emerging Trends in Engineering, Sciences and Technology (ICES&T) (pp. 1-6). IEEE.
18. Ibrahim, M., Bajwa, I. S., Sarwar, N., Hajjej, F., & Sakr, H. A. (2023). An Intelligent Hybrid Neural Collaborative Filtering Approach for True Recommendations. IEEE Access.

19. Chavan, Jaymala, and Priyanka Desai. "Data leakage detection using data allocation strategies." International Journal of Advances in Engineering & Technology 6.5 (2013): 2033.
20. Costante, Elisa, et al. "A hybrid framework for data loss prevention and detection." 2016 IEEE Security and Privacy Workshops (SPW). IEEE, 2016.
21. Naeem, Awad Bin, et al. "Deep Learning Models for Cotton Leaf Disease Detection with VGG-16." International Journal of Intelligent Systems and Applications in Engineering 11.2 (2023): 550-556.
22. A. A. Eladl, M. I. El-Afifi, M. A. Saeed, & M. M. El-Saadawi, Optimal operation of energy hubs integrated with renewable energy sources and storage devices considering CO2 emissions. International Journal of Electrical Power & Energy Systems, 2020,117, 105719.
23. A. A. Eladl, M. I. El-Afifi, M. M. El-Saadawi, & B. E. Sedhom, A review on energy hubs: Models, methods, classification, applications, and future trends. Alexandria Engineering Journal, 2023, 68, 315-342.
24. M. I. El-Afifi, M.M. Saadawi, & A. A. Eladl, Cogeneration Systems Performance Analysis as a Sustainable Clean Energy and Water Source Based on Energy Hubs Using the Archimedes Optimization Algorithm. Sustainability, 2022,14(22), 14766.
25. A. A. Eladl, A. A., M. E. El-Afifi, & M. M. El-Saadawi, Communication technologies requirement for energy hubs: a survey. In 2019 21st International Middle East Power Systems Conference (MEPCON), 2019, (pp. 821-827).
26. A. A. Eladl, A. A., M. E. El-Afifi, & M. M. El-Saadawi, Optimal power dispatch of multiple energy sources in energy hubs. ERJ. Engineering Research Journal, 2018, 41(4), 279-287.
27. A. A. Eladl, M. I. El-Afifi, M. M. El-Saadawi, & B. E. Sedhom, Distributed optimal dispatch of smart multi-agent energy hubs based on consensus algorithm considering lossy communication network and uncertainty. CSEE Journal of Power and Energy Systems., 2023.
28. Mansour, Nehal A., et al. "Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy." Journal of ambient intelligence and humanized computing (2022): 1-33.
29. Rabie, A. H., Mansour, N. A., Al-Husseiny, A., & Saleh, A. I. (2021). A Review on COVID-19 Patients Detection Using Data Mining and IoT Technology. Nile Journal of Communication and Computer Science, 1(1), 19-28.
30. Mansour, Nehal A., et al. "The Role of IoT in COVID-19." Nile Journal of Communication and Computer Science 1.1 (2021): 29-38.
31. Rabie, Asmaa H., et al. "Expecting individuals' body reaction to Covid-19 based on statistical Naïve Bayes technique." Pattern Recognition 128 (2022): 108693.
32. Rabie, Asmaa H., Ahmed I. Saleh, and Nehal A. Mansour. "A Covid-19's integrated herd immunity (CIHI) based on classifying people vulnerability." Computers in Biology and Medicine 140 (2022): 105112.
33. H. A. Sakr et al., "Al-based Traffic System: A Novel Approach," 2023 24th International Middle East Power System Conference (MEPCON), Mansoura, Egypt, 2023, pp. 1-6, doi: 10.1109/MEPCON58725.2023.10462361.
34. Khan, S. H., Alahmadi, T. J., Alsahfi, T., Alsadhan, A. A., Mazroa, A. A., Alkahtani, H. K., ... & Sakr, H. A. (2023). COVID-19 infection analysis framework using novel boosted CNNs and radiological images. Scientific Reports, 13(1), 21837.

35. M. I. El-Afifi, H. A. Sakr, Security Issues and Challenges for IoT-based Smart Multi Energy Carrier Systems. Nile Journal of Communication and Computer Science, 2023.
36. H. A. Sakr., M. I. El-Afifi, Mechanisms of system penetration. Nile Journal of Communication and Computer Science, 2023.