

A COMPREHENSIVE APPROACH TO PREVENTING DATA LEAKAGE AND STRENGTHENING CYBERSECURITY

TMP-2023-24-082

PROPOSAL PROJECT REPORT

AHAMED RR – IT20650902

B.Sc. (Hons) in Information Technology Specializing in Cyber Security

Department of Computer System and Engineering

Sri Lanka Institute of Information Technology

August 2023

A COMPREHENSIVE APPROACH TO PREVENTING DATA LEAKAGE AND STRENGTHENING CYBERSECURITY

TMP-2023-24-082

PROPOSAL PROJECT REPORT

SAFEGUARDING SYSTEMS BY IDENTIFYING UNUSUAL USER PATTERNS

AHAMED RR – IT20650902

Supervisor – Mr. Amila Senarathne

Co – Supervisor –

B.Sc. (Hons) in Information Technology Specializing in Cyber Security

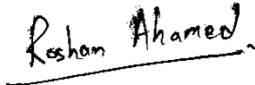
Department of Computer System and Engineering

Sri Lanka Institute of Information Technology

August 2023

DECLARATION

We declare that this is our own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Name	Student ID	Signature	Date
Ahamed RR	IT20650902		08.25.2023

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.

Signature of the Supervisor

Date

.....

.....

Signature of the Co-Supervisor

Date

.....

.....

ABSTRACT

Nowadays, the escalating threat of data breaches has become a paramount concern for businesses. The safeguarding of sensitive data is a top priority, necessitating the attention of top management, IT administrators, and experts alike. Traditional security measures like firewalls are proving inadequate in the face of evolving cyber threats. Data Loss Prevention (DLP) systems are a new desire in the struggle for data security. This research initiative comprises 4 interconnected subcomponents aimed at providing a comprehensive strategy to minimize data loss and enhance cybersecurity.

The first subcomponent, "Safeguarding Systems by Identifying Unusual User Patterns " uses a machine learning algorithm to proactively identify unusual user behavior and access patterns, enabling timely responses to potential threats. The second, "Utilizing NLP Techniques for Enhanced Data Protection" explores natural language processing algorithms to automatically identify sensitive information in text, boosting data safety by detecting insider risks. The third, "Unveiling Patterns and Anomalies to Mitigate Data Breach Risks" employs innovative data analysis tools and machine learning to uncover hidden patterns and minimize breach risks. Lastly, "Malicious Image Detection and Classification Using Deep Learning Techniques" focuses on defending against image-based cyberattacks by utilizing convolutional neural networks to distinguish between legitimate and malicious images.

In this research my sub research component is Safeguarding Systems by Identifying Unusual User Patterns. A traditional rule-based approach or a single algorithm approach has limitations when it comes to capturing complex patterns of anomalous user behavior and adapting to evolving threats. As a solution in this research, we aim to develop a tool that can effectively identify unusual user patterns by using multiple anomaly detection algorithms. Using ensemble methods, this research aims to advance cybersecurity practices by proactively detecting cyber threats and strengthening digital security.

TABLE OF CONTENTS

DECLARATION.....	3
ABSTRACT.....	4
TABLE OF CONTENTS	5
1. INTRODUCTION	6
1.1 Background & Literature Survey.....	6
1.2 Research Gap	8
1.2.1 Research Gap Comparison Chart	9
1.3 Research Problem	10
2. OBJECTIVES.....	11
2.1 Main Objective	11
2.2 Sub Objective.....	11
3. METHODOLOGY	12
3.1 System Diagram.....	13
3.3 Technologies	13
4. PROJECT REQUIREMENTS.....	14
4.1 Functional Requirements	14
4.2 Non-Functional Requirements.....	14
4.3 Software Requirements.....	14
5. WORK BREAKDOWN STRUCTURE (WBS)	15
6. GANTT CHART.....	16
7. BUDGET	17
8. COMMERCIALIZATION	18
9. REFERENCES	19
APPENDICES	20

1. INTRODUCTION

1.1 Background & Literature Survey

Why safeguarding systems is important –

The security of systems is critical in today's digital world. The fast advancement of technology and the internet has resulted in an increased threat of cyber-attacks. The attacks could involve anything from unauthorized access to sensitive data to modern data breaches. These security breaches can be dangerous resulting in financial losses, reputational harm, and legal issues for organizations. Furthermore, the growth of remote work and the clouds has increased the attack surface, making system security even more vital. As a result, safeguarding systems have become a necessary approach for organizations that want to protect their assets, sensitive information, and stakeholder confidence [1].

What are unusual user patterns –

Unusual user patterns in a system are variations from normal or expected user behavior. These variations may be suggestive of possible security vulnerabilities, data breaches, or even insider attacks. Users accessing sensitive data outside of regular working hours, multiple attempts to log in with wrong passwords, or a rapid rise in data transmission to external devices are examples of odd usage habits. Identifying these patterns is critical for taking preventative action since they might indicate possible security weaknesses [2].

How to identify unusual user patterns –

Organizations take advantage of different kinds of strategies for identifying unusual user patterns, including:

1. **Statistical Analysis:** Statistical analysis is the process of identifying patterns that differ considerably from the norm by employing mathematical and statistical models. Tracking login timings and comparing them to previous data, for example, could identify anomalous login patterns of user behavior.
2. **Machine Learning:** Machine learning algorithms are increasingly being used to spot anomalies in user behavior data. These algorithms are capable of analyzing large datasets and detecting deviations from pre-established user behavior models. They were excellent in detecting complicated and frequently changing patterns, allowing organizations to take an active approach to security.
3. **Pattern Recognition:** The automatic identification of patterns within data is a key component of pattern recognition techniques, which are frequently combined with machine learning. This may involve identifying behavioral patterns that could point to unusual user behaviors.

4. User Profiling: Developing user profiles that outline normal user behavior could help in spotting departures from accepted standards. Any behavior that deviates from these patterns might be marked as potentially not usual.

The primary area of study for intrusion detection systems nowadays is anomaly detection; its main feature is the ability to identify unknown attack mechanisms through the detection of anomalous user behavior. Establishing common usage patterns and figuring out how to take advantage of them to compare and evaluate the present user behavior are the main challenges in anomaly detection [3]. And in the earlier stages, Access control policies were typically employed in the user behavioral regulation to carefully restrict users' behavior rights, but they were unable to stop the active destruction of legitimate users [4]. The deployment of user behavior anomaly detection approaches can successfully stop authorized users from engaging in active malicious activity. User actions frequently reveal historical trends. for that, by developing user behavioral patterns to prevent unusual user patterns, identifying defects between patterns, and using anomaly detection of user behaviors [5]. But at the moment, the following are user behavior anomaly detection systems' shortcomings: inadequate capacity for processing huge numbers of data automatically [6].

And another research demonstrated that a pattern mining-based user-behavior anomaly detection algorithm can automatically process large user behavior audit data; to a certain degree, the accuracy of detection could be increased [7]. The literature highlights the increasing demand for cybersecurity protection of systems through anomaly detection. Ensemble approaches have the ability to dramatically increase the accuracy and effectiveness of anomaly detection in user behavior, which would improve security measures. This is especially true when combined with modern feature extraction methods.

1.2 Research Gap

Protecting systems against cyber-attacks and data breaches is a crucial issue for organizations in the dynamic field of cybersecurity. Unusual user behaviors that may indicate security vulnerabilities or breaches are crucially identified by anomaly detection, a key element of cybersecurity. While major advancements in the creation of anomaly detection systems, a significant research gap still exists in the field. This gap is specifically focused on the scant investigation of merging different anomaly detection systems to improve the precision of forecasts of atypical user behaviors.

- The Dominance of Rule-Based Approaches –

Many anomaly detection systems have historically used rule-based techniques. These systems are designed to spot violations of established norms or restrictions. Their rigidity presents problems when dealing with complex and developing cyber threats, even if they can successfully catch known abnormalities. Rule-based systems may find it difficult to keep up with the constant evolution of cyber attackers' tactics, methods, and procedures, which may lead to false negatives or missing threats [8].

- Single Algorithm-Centric Approaches –

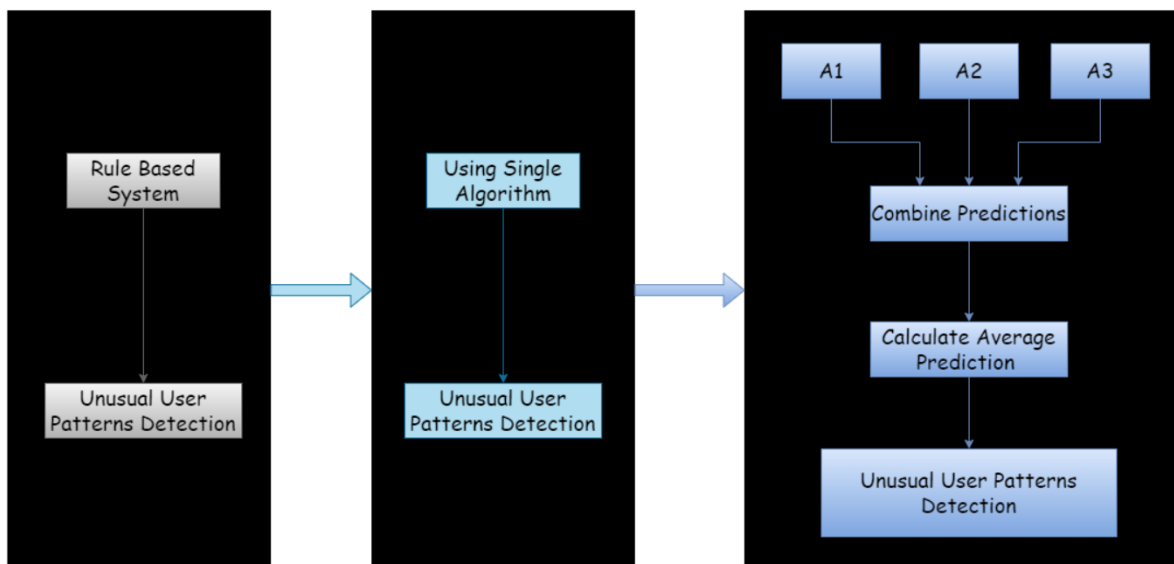
The growing number of single algorithm-centric techniques is at the core of the research gap. Using a single algorithm to identify unexpected user patterns is the focus of a significant percentage of current anomaly detection research efforts. While these algorithms may perform well under some circumstances, they have limits when dealing with a variety of threats that are constantly changing. Single-algorithm techniques may not be able to fully capture complex, nuanced patterns that point to security issues. The growing complexity of cyber-attacks only makes this constraint stronger [9].

- The Need for Ensemble Methods in Anomaly Detection –

The research gap indicates that ensemble technique research in anomaly identification is of critical importance. By combining the results of several anomaly detection algorithms to provide final anomaly predictions, ensemble techniques provide a strong solution. This strategy makes use of the combined knowledge of many algorithms and could improve the system's overall detection performance, resilience, and flexibility.

1.2.1 Research Gap Comparison Chart

Aspect	Rule-Based Methods	Single Algorithms	Combining Multiple Algorithms
Detection Approach	Rule-based approach, using pre-established restrictions and patterns.	single anomaly detection algorithm to identify unusual user patterns.	The approach focuses on the combination of multiple anomaly detection algorithms.
Detection Performance	High false positives	The performance depends on the algorithm selected. Some false positive or false negative issues.	Potential for improved accuracy, reducing false positives.
Anomaly Detection Effectiveness	Less effective in identifying complex.	Effectiveness is based on selected algorithm.	More effective in identifying a wider range of anomalies.



1.3 Research Problem

The main research problem being addressed pertains to the need to implement safety mechanisms that can effectively detect and identify unusual user patterns. Organizations need methods and techniques to analyze user behavior and access patterns to detect suspicious activities that may indicate unauthorized access or malicious intent, or potential data leaks. However, maintaining an accurate balance between false positives and false negatives poses a significant difficulty in the complicated area of anomaly detection in user patterns.

1. The Need for Unusual User Pattern Identification

In the field of cybersecurity, protecting systems against unauthorized access, data breaches, and security risks is critical. Identifying unusual user patterns that might indicate possible security issues is a critical component of this approach. It is without a doubt necessary for safeguarding systems to efficiently recognize unusual user patterns. These patterns include a wide range of actions, from suspicious login attempts and unusual login schedules to illegal data access and abnormal session durations. To identify threats early, reduce risks, and maintain data integrity, it is essential to recognize these user patterns.

2. Balancing False Positives and False Negatives

The particular balance between false positives and false negatives in anomaly identification is a key difficulty. When the system wrongly identifies typical user behavior as suspicious or unusual, false positives happen. A torrent of unnecessary alarms, resource waste, and stress on staff members charged with reviewing these alerts might result from this. False negatives, on the other hand, happen when actual security risks or atypical usage habits go unnoticed. These undiscovered abnormalities have the potential to cause serious security breaches, data leaks, and monetary losses.

By merging several anomaly detection algorithms and averaging their findings to get the final anomaly predictions on user behaviors, our study intends to tackle this enormous difficulty. In doing so, we want to strengthen security systems, minimize the need for alerts, and lower the likelihood of security breaches, all of which will help develop cybersecurity practices and ensure the safety of organizational assets and sensitive data.

2. OBJECTIVES

2.1 Main Objective

This research's my primary objective is to develop, create, and implement an innovative tool that uses machine learning algorithms to accurately detect and describe unusual user patterns by examining user behavior. Developing this kind of tool is of most significance at a time of growing cyber threats and complex digital ecosystems.

Our main goal is to develop a reliable and flexible way that can accurately identify departures from established user norms. In order to process and evaluate the huge number of user activity data, this tool will employ powerful machine learning methods. As a result, it will be able to spot even small anomalies that could escape detection of rule-based systems in the past.

2.2 Sub Objective

- Combine Multiple Anomaly Detection Algorithms –

We understand that using a single algorithm method for detecting anomalies would not be enough to completely capture the wide range of unusual user behaviors. So, this research focuses on the combination of several anomaly detection algorithms to mitigate the impact of this limitation.

- Aggregate Results for Final Predictions –

We propose combining the outputs of these many algorithms for getting at the final abnormal predictions on user patterns instead of depending just on the output of a single algorithm. By using the combined predictions of many algorithms, this ensemble technique improves the reliability and accuracy of detection.

- Enhanced Accuracy –

Combining multiple algorithms can potentially improve the accuracy of detecting unusual user patterns. In combination, the unique perspectives, strengths, and assumptions of each algorithm contribute to more accurate anomaly detection.

3. METHODOLOGY

In this section, we will detail the systematic methodology for designing a tool focused on safeguarding the system by finding unusual user patterns in user activity data, with an extra focus on the use of ensemble approaches.

- Data Collection and Preprocessing –

Data associated with user behavior may be gathered from a variety of sources, including system logs, application records, and network traffic. The data set represents normal user patterns and unusual user patterns in the target environment. In the preprocessing part, we are going to ensure consistency and quality, clean and preprocess the obtained data, and handle missing values. As per the requirement for additional analysis, we are going to normalize the dataset.

- Algorithm Selection and Implementation –

In this step, anomaly detection algorithms are chosen, implemented, and evaluated for user behavior data suitability, considering elements like interpretability, computational effectiveness, and adaptability to changing user patterns. After the selection of the algorithms, apply the selected anomaly detection methods. Make sure they are set up to efficiently handle user behavior data. This stage involves fine-tuning and optimizing the parameters for each algorithm.

- Ensemble Method Development –

In this phase, we are going to choose several ensemble methods to combine multiple anomaly detection algorithms with the aim of improving anomaly detection accuracy. Embed the chosen ensemble techniques inside the framework of the tool's design. Finally, develop strategies for integrating the findings of different anomaly detection algorithms inside the ensemble.

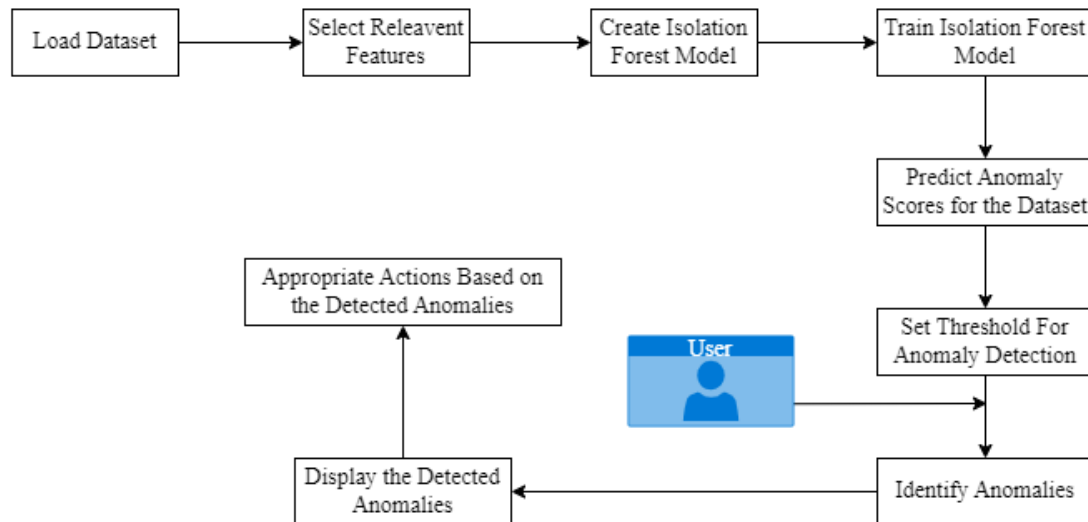
- Model Training and Evaluation –

A part of the preprocessed data needs to be set aside for training the anomaly detection models. and to evaluate the performance of the models, use cross-validation methods. next, Divide the dataset into training and validation sets, train the models, and analyze their performance using measures such as accuracy and precision.

- Performance Testing and Validation –

After completion of the development, we can test the tool with real-world user behavior data obtained from various situations and organizations. Next, we can validate the effectiveness of the tool in identifying unusual user patterns and measuring performance.

3.1 System Diagram



3.3 Technologies

- Machine Learning
- Flask API
- Python

4. PROJECT REQUIREMENTS

4.1 Functional Requirements

To ensure the tool performs its primary task of identifying unusual user patterns and potential security threats within user behavior data we need pieces of information like

- Login attempts within a period.
- Session durations.
- Login frequencies.

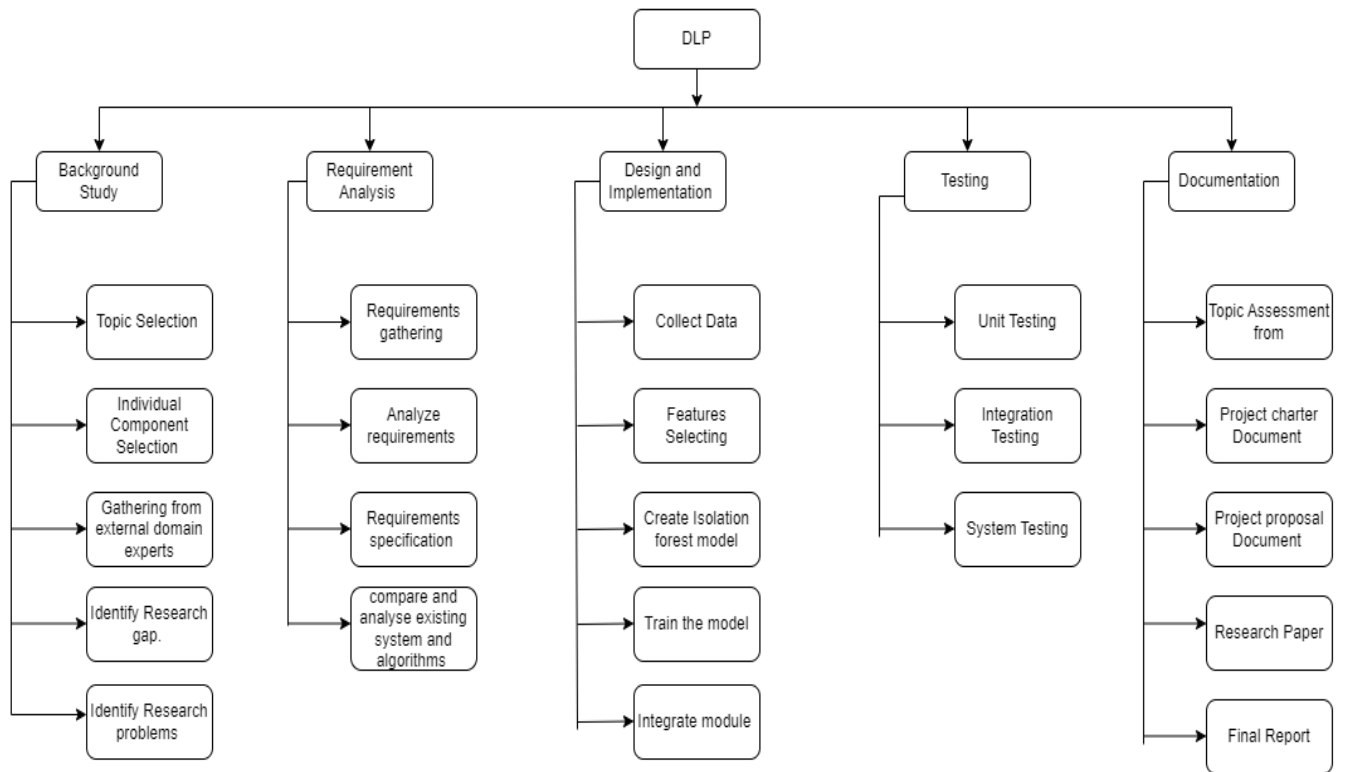
4.2 Non-Functional Requirements

- Reliability
- High performance in anomaly detection.
- Usability with the user-friendly interface.

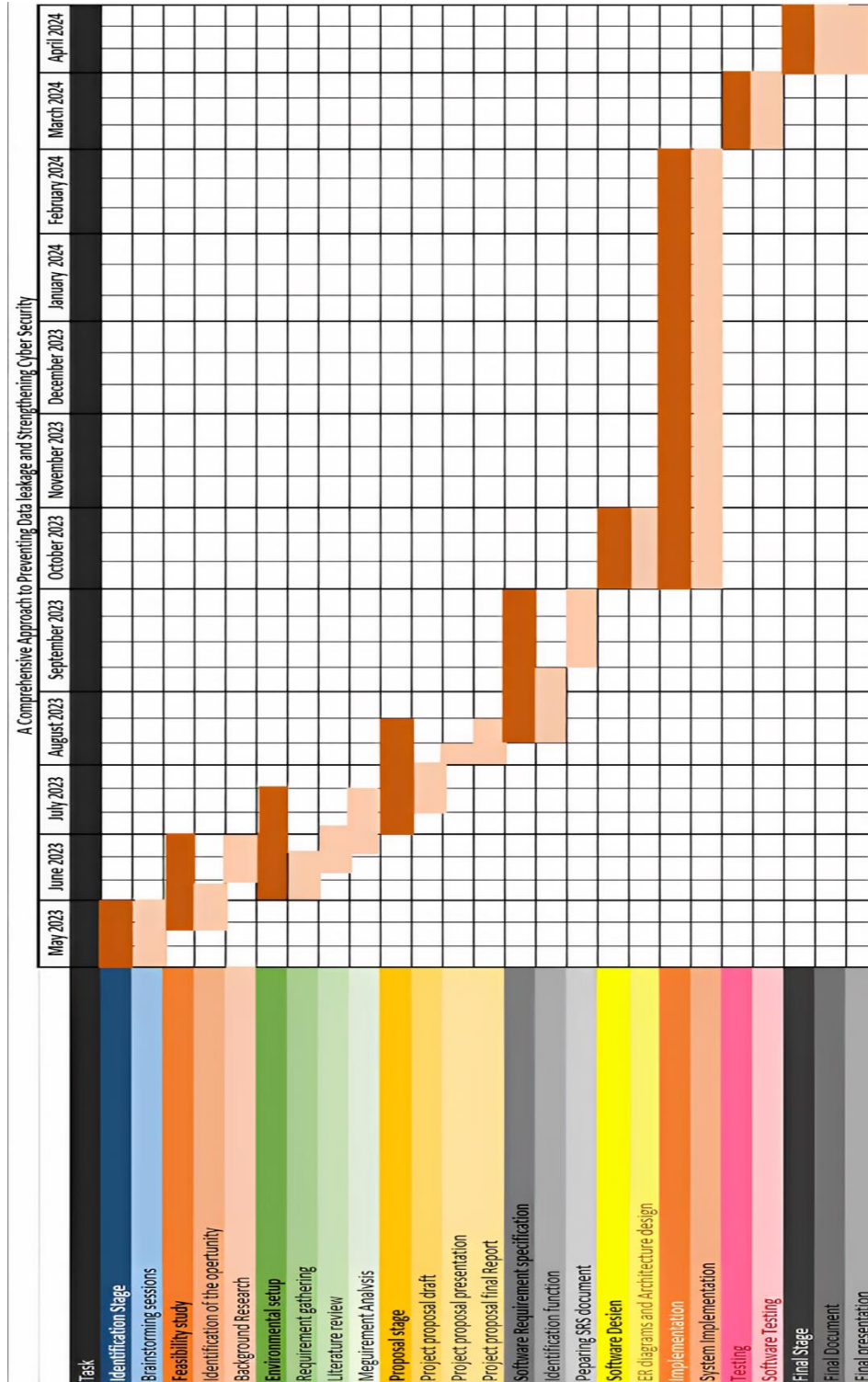
4.3 Software Requirements

- PyCharm

5. WORK BREAKDOWN STRUCTURE (WBS)



6. GANTT CHART



7. BUDGET

Resources	Price(LKR)
Internet	2000.00
Stationary Materials	1000.00
Electricity	2000.00
Hardware Equipment	3000.00
Paper publish cost	5000.00
Training & Testing cost	4000.00
Total	17000.00

8. COMMERCIALIZATION

- **Create a sales plan for the industry.**

Industry analysis: Conduct thorough market research to identify industries with the highest demand for comprehensive cybersecurity solutions. Understand their pain points, regulatory requirements, and specific challenges associated with data leakage and cybersecurity.

Tailored fee Proposition: Craft a compelling value proposition that directly addresses the unique needs of each targeted industry. Highlight how our system's capabilities align with its requirements, showcasing its effectiveness in preventing data leakage and strengthening overall cybersecurity.

Strategic Partnerships: Forge partnerships with influential industry associations, organizations, and thought leaders. Collaborate to co-host webinars, workshops, or events that position our system as an innovative solution, gaining credibility and expanding our reach within the industry.

- **Design customer subscription plans.**

Tiered Plans: Develop various subscription levels, each catering to different business sizes and needs. Offer options like basic, standard, and premium plans, each with a distinct set of features and capabilities.

Scalability: Ensure that your subscription plans are designed to accommodate the growth and changing requirements of businesses. Provide flexibility for clients to upgrade or adjust their plans as their needs evolve.

Customization Flexibility: Integrate customization options within subscription plans, allowing clients to tailor features based on their specific data protection needs. This ensures that they only pay for the functionalities they require.

- **Provide excellent customer support.**

Dedicated support team: Assign a team of knowledgeable support representatives to promptly address inquiries and issues.

Multi-Channel assist: Provide support through various channels such as email, live chat, and phone.

24/7 Availability: offer round-the-clock assistance for critical concerns and urgent inquiries.

Knowledge Base: Develop an online resource with FAQs, tutorials, and troubleshooting guides.

Continuous training: Offer training sessions to help clients maximize the benefits of the system.

9. REFERENCES

- [1] Keshk, M., Turnbull, B., Sitnikova, E., Vatsalan, D., & Moustafa, N., "Privacy-preserving schemes for safeguarding heterogeneous data sources in cyber-physical systems.," in *Practical Innovations*, 2021.
- [2] Davide Fauri, Sandro Etalle, Jerry den Hartog, Nicola Zannone, "A Hybrid Framework for Data Loss Prevention and Detection," in *ymposium on Security and Privacy Wor*, 2016.
- [3] Yang Yang, Juan Hao, Jianguang Zhao, Cihang Chen, Haoyue Sun, "Computer User Behavior Anomaly Detection Based on," in *Wiley Hindawi, china*, 2022.
- [4] C. Wu, R. Yu, B. Yan et al, "Data design and analysis based on," in *Journal of Intelligent and Fuzzy Systems*, 2020.
- [5] Y. Zhou, R. Xie, T. Zhang, and J. Holguin-Veras, "Joint Distribution Center Location Problem for Restaurant Industry Based on Improved K-means Algorithm with Penalty," in *IEEE Access*, 2020.
- [6] M. Veeraiyan, S. Kousika, J. Senthilkumar, J. C. Antonysami, "An Optimized Mobile Cloud Computational Offloading," in *Journal of Computer Science*, 2019.
- [7] P. T. Seta and K. D. Hartomo, "Mapping Land Suitability for Sugar Cane Production Using K-means Algorithm with Leaflets Library to Support Food Sovereignty in Central Java," in *Khazanah Informatika Jurnal Ilmu Komputer dan Informatika*, 2020.
- [8] Manu Bijone, "A Survey on Secure Network: Intrusion," in *American Journal of Information Systems*, MP, 2016.
- [9] The Dominance of Rule-Based Approaches, WILLIAM HURST, MICHAEL MACKAY, ABDENNOUR EL RHALIBI, "Density-Based Outlier Detection for Safeguarding Electronic Patient Record Systems," in *IEEE Access*, Liverpool, 2019.

APPENDICES

ahamed

ORIGINALITY REPORT

13%

SIMILARITY INDEX

11%

INTERNET SOURCES

4%

PUBLICATIONS

9%

STUDENT PAPERS

PRIMARY SOURCES

1

www.coursehero.com

Internet Source

4%

2

www.hindawi.com

Internet Source

2%

3

digital.lib.usu.edu

Internet Source

1%

4

Abebe Diro, Shahriar Kaiser, Athanasios V. Vasilakos, Adnan Anwar, Araz Nasirian, Gaddisa Olani. "Anomaly Detection for Space Information Networks: A Survey of Challenges, Schemes, and Recommendations", Institute of Electrical and Electronics Engineers (IEEE), 2023

Publication

1%

5

digilib.k.utb.cz

Internet Source

1%

6

dblp.dagstuhl.de

Internet Source

1%

7

arimaa.com

Internet Source

1%