**“DATA EXFILTRATION DETECTION THROUGH EMAIL-BASED INSIDER THREATS”**

***A Report submitted***

***in partial fulfilment for the Degree of***

**Masters of Science**

**IN**

**Cyber Security**

***Submitted By***

**Disha Sharma**

**(240103002014)**

***Under the Supervision of***

**Mr. Dharmesh Dave**

**(Assistant Professor)**

***Submitted to***

**SCHOOL OF CYBER SECURITY & DIGITAL FORENSICS,**

**NATIONAL FORENSIC SCIENCES UNIVERSITY**

**GANDHINAGAR – 382007, GUJARAT, INDIA.**

**MAY, 2025**

DECLARATION

I **DISHA SHARMA** having Enrolment Number **240103002014** hereby declare that

a. The work contained in the dissertation report entitled **“DATA EXFILTRATION DETECTION THROUGH EMAIL-BASED INSIDER THREATS”** is being submitted in partial fulfilment for the award of the degree of **M.Sc. Cyber Security** to **School of Cyber Security & Digital Forensics** is an authentic record of my own work done under the supervision of **Mr. Dharmesh Dave**.

b. The work has not been submitted to any other Institute/ School / University for any degree or diploma.

c. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the School.

d. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the dissertation and giving their details in the references.

e. Whenever I have quoted written materials from other sources and due credit is given to the sources by citing them.

f. From the plagiarism test, it is found that the similarity index of whole dissertation is less than 10 % as per the university guidelines.

**Date:**

**Place:**

**Signature of Student**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Signature & Date**

**Supervisor**

CERTIFICATE

This is to certify that the work contained in the dissertation entitled “DATA EXFILTRATION DETECTION THROUGH EMAIL-BASED INSIDER THREATS”, submitted by Disha Sharma (Enrol. No.: 240103002014) in partial fulfilment of the requirement for the award of the degree of M.Sc. Cyber Security to the National Forensic Sciences University, Gandhinagar, Gujarat is a record of Bonafide work carried out by him/her under the direct supervision and guidance of Mr. Dharmesh Dave.

**Date:**

**Place:**

**Supervised By:**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Name & sign of the Supervisor 1) (Name & sign of the Supervisor 2, if any)

Designation,

School of Cyber Security and Digital Forensics,

National Forensic Sciences University,

Gandhinagar, India, 382421

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dean, School of Cyber Security and Digital Forensics,

National Forensic Sciences University,

Gandhinagar, India, 382421

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and deepest appreciation to **Mr. Dharmesh Dave**, my esteemed mentor, for his exceptional guidance and unwavering support during the course of this dissertation.

His invaluable insights and meticulous attention to detail have been instrumental in shaping and refining this research project. His exceptional guidance, profound expertise and constructive feedback have been the stepping stones of my intellectual and academic journey and his dedication has been commendable throughout.

**Mr. Dharmesh Dave’s** mentorship has not only enhanced my understanding of the subject matter but also instilled in me a deep sense of discipline and scholarly rigour. His unwavering commitment to excellence has served as a constant source of inspiration, urging me to push the boundaries of knowledge in the field of Threat Intelligence and Intrusion Detection.

To **Mr. Dharmesh Dave** and all my lab mates who have contributed to this dissertation in various capacities, I offer my profound gratitude. Your exceptional guidance, support and expertise have not only impressed me but also served as the cornerstone of this research endeavour.

With Sincere Regards,

**Disha Sharma**

**MSc Cyber Security**

ABSTRACT

*This dissertation introduces the development and design of a real-time detection tool targeting the detection of data exfiltration attempts through email-based insider threats in a small to mid-sized organizational environment. The key aim is to detect accidental as well as malicious insider behaviour—such as phishing email interactions, misdirected emails, and unauthorized transmission of confidential information—through a mix of static rules, behavioural detection, and content-aware methods.*

*The tool connects to the Gmail API to track incoming and outgoing email activity, providing support for real-time alerting through push notifications. Confidential corporate files are labeled by sensitivity level, and a two-layered detection system is put in place using fuzzy hashing and NLP-content-based verification to detect attempts at circumventing detection through file alteration or coded messaging. A built-in alerting system classifies threats by severity, with color-coded CLI logs and optional notifications, providing greater visibility to security teams.*

*The dissertation is organized into several phases, starting with the identification of the problem space and the setting of the objectives; describing the system architecture and major components; describing the detection methods, rule-based and anomaly detection models; describing the implementation process and integration with Gmail; testing the performance of the system using simulated insider scenarios. The last phase concludes the results and provides suggestions for future improvement, including SIEM integration and adaptive learning mechanisms.*

*The results verify the tool's success in identifying varied types of data exfiltration via email and its usability in a 15–20 employee setup. The project accomplishes the fundamental aim of providing improved insider threat visibility with low false positives and low operational overhead.*

**Keywords: Real-time detection, Insider behaviour, Gmail API, CLI logs, Rule-based, SIEM Integration, Data Exfiltration, False Positives, Severity, Built-in alerting system**

LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| API | Application Programming Interface |
| AUP | Acceptable Usage Policy |
| BYOD | Bring Your Own Device |
| CLI | Command Line Interface |
| IDS | Intrusion Detection System |

|  |  |
| --- | --- |
| IBM | International Business Machines Corporation |
| IT | Information Technology |

|  |  |
| --- | --- |
| NLP | Natural Language Processing |
| RBAC | Role Based Access Control |

|  |  |
| --- | --- |
| SIEM | Security Information and Event Management |
| ZIP | Zipped |

LIST OF TABLES

[Table 1: Comparative Analysis 4](#_Toc197605876)

[Table 2 : Evaluation Metrics 13](#_Toc197605877)

[Table 3: Evaluation Results 13](#_Toc197605878)

[Table 4: Test Cases. 15](#_Toc197605879)

LIST OF FIGURES

[Figure 1: Test Users 11](#_Toc197605843)

[Figure 2: OAuth Screen 11](#_Toc197605844)

[Figure 3: Token Credentials 12](#_Toc197605845)

[Figure 4: Zerotrace – AUP 12](#_Toc197605846)

[Figure 5: Confidential Attachments 12](#_Toc197605847)

[Figure 6: Phishing Email 18](#_Toc197605848)

[Figure 7: Result 18](#_Toc197605849)

[Figure 8: Confidential Attachment 19](#_Toc197605850)

[Figure 9: Result 19](#_Toc197605851)

[Figure 10: Personal Email Upload 20](#_Toc197605852)

[Figure 11: Result 20](#_Toc197605853)

[Figure 12 : Modified Confidential Attachment 21](#_Toc197605854)

[Figure 13: Result 21](#_Toc197605855)

[Figure 14 : Google Drive Link as Attachment 22](#_Toc197605856)

[Figure 15: Result 22](#_Toc197605857)

[Figure 16: Result 23](#_Toc197605858)

[Figure 17: Obfuscated Email 23](#_Toc197605859)

[Figure 18: Normal Email 24](#_Toc197605860)

[Figure 19: Result 24](#_Toc197605861)

[Figure 20: Personal Email Upload 25](#_Toc197605862)

[Figure 21: Result 25](#_Toc197605863)

[Figure 22: Confidential Attachment 26](#_Toc197605864)

[Figure 23: Result 26](#_Toc197605865)

[Figure 24: Obfuscated Email 27](#_Toc197605866)

[Figure 25: Result 27](#_Toc197605867)

[Figure 26 28](#_Toc197605868)

[Figure 27: Confidential Attachment 28](#_Toc197605869)

[Figure 28: Result 1 28](#_Toc197605870)

[Figure 29: Result 2 28](#_Toc197605871)

[Figure 30: Progress Report 33](#_Toc197605872)

[Figure 31: Plagiarism Report 34](#_Toc197605873)

TABLE OF CONTENTS

[DECLARATION II](#_Toc197605944)

[CERTIFICATE III](#_Toc197605945)

[ACKNOWLEDGEMENTS IV](#_Toc197605946)

[ABSTRACT V](#_Toc197605947)

[LIST OF ABBREVIATIONS VI](#_Toc197605948)

[LIST OF TABLES VII](#_Toc197605949)

[LIST OF FIGURES VIII](#_Toc197605950)

[TABLE OF CONTENTS IX](#_Toc197605951)

[1. INTRODUCTION 1](#_Toc197605952)

[1.1 Motivation 1](#_Toc197605953)

[1.2 Scope of Project 2](#_Toc197605954)

[1.3 Organization of report 3](#_Toc197605955)

[2. THEORETICAL BACKGROUND AND LITERATURE SURVEY 4](#_Toc197605956)

[2.1 Comparative Analysis of Existing Schemes 4](#_Toc197605957)

[2.2 Research Findings 5](#_Toc197605958)

[3. THE PROPOSED MODEL AND IMPLEMENTATION STRATEGY 7](#_Toc197605959)

[3.1 The Problem Statement (Revisited) 7](#_Toc197605960)

[3.2 The System Model 7](#_Toc197605961)

[3.3 Methodology 9](#_Toc197605962)

[*1. Email Monitoring* 9](#_Toc197605963)

[*2. Static Rule-Based Threat Detection* 9](#_Toc197605964)

[*3. File Confidentiality Tagging* 9](#_Toc197605965)

[*4. File Tagging and Content Matching* 9](#_Toc197605966)

[*5. NLP-Based Content Analysis* 10](#_Toc197605967)

[*6. Behavioural Analysis and Threat Classification* 10](#_Toc197605968)

[*7. Logging and Real-Time Alerting* 10](#_Toc197605969)

[4. THEORETICAL AND EMPIRICAL RESULT ANALYSIS 11](#_Toc197605970)

[4.1 Experimental Setup 11](#_Toc197605971)

[4.2 Evaluation Metrics 13](#_Toc197605972)

[4.3 Test Application/ Scenario 15](#_Toc197605973)

[4.4 Theoretical Analysis 16](#_Toc197605974)

[4.5 Implementation Result Analysis 18](#_Toc197605975)

[5. CONCLUSION 29](#_Toc197605984)

[5.1 Final Thoughts 29](#_Toc197605985)

[6. FUTURE SCOPE 30](#_Toc197605986)

[7. BIBLIOGRAPHY 31](#_Toc197605987)

[PROGRESS REPORT 33](#_Toc197605988)

[PLAGIARISM REPORT 34](#_Toc197605989)

1. INTRODUCTION
   1. Motivation

The past few years have seen a surge in the number of cyber-attacks and cyber threats, globally. Of these legitimate and accounted threats, 83% of organizations reported atleast one insider threat in 2024 according to a study conducted by IBM. An even more surprising fact is that insider threat-based attacks saw an increase of about five times than those accounted in 2023. While these threats are on a rise, it is important for organizations to address the risks posed by them in the digital ecosystem while putting into practice effective threat management strategies to address them.

Several studies were conducted to identify the root causes of insider threat-based attacks and the following four reasons were prominent in facilitating insider threat attacks:

A) Complicated IT environments – Many organizations facilitate work-from-home environments and BYOD facilities even more after the COVID pandemic. This has increased the potential attack surface of organizations and provided greater access to confidential attachments of the organizations on employees’ remote devices.

B) Inadequate Security Measures – Several small and medium scaled organizations are not updated with latest security measures and do not keep devices patched in accordance with recent security vulnerabilities. This proves to be a great risk for the organizations in the long run.

C) Lack of Employee Training and Awareness – Employees of the organization are not made familiar with acceptable use policies and user awareness protocols during their initiation that makes them vulnerable to external threat actors such as phishing emails.

D) Weak Enforcement Policies - Although 93% of respondents in the report said that strict visibility and control was an important factor for them, only 36% actually had an effective solution in place for unified visibility and access control.

According to Teramind Inc. about 67% of malicious insiders are likely to email sensitive data to outside parties leading to data exfiltration and major incident costs for the organizations. There could be several motives for insiders to facilitate such attacks, for instance – money, disgruntling experiences, revenge or even accidental emailing.

There are tools in the market to detect precursors and indicators regarding potential data exfiltration attacks, but some of them do not acount for false positives heavily while others are not able to classify the intention of the employee as malicious or accidental.

Thus, in developing this research project, my aim was to detect indicators of compromise for a data exfiltration attack through insiders of the organization via email communications and classify the attack intention as “malicious” or “accidental”.

* 1. Scope of Project

The following are the objectives that the project aims to cover:

A) **Phishing Detection** – To detect phishing attempts either from senders external to the organization or within the organization and to flag them with a relevant severity level and intention classification. To also detect spoofed emails and hyperlink attachments by examining the body of email messages using pre-defined phishing keywords. To also flag repeated phishing attempts from insiders or outsiders.

B) **Confidential Communication** – To detect emails containing confidential keywords being sent to external receivers not a part of the organization and to flag these based on severity levels, intention classification and repeated attempts.

C) **Confidential Attachments** – To sign confidential files and detect emails containing these confidential attachments addressed to employees within the organization not having access to them or external unauthorized entities. To flag these alerts based on severity levels, intention classification and repeated attempts.

D) **ZIP Attachments** – To flag emails containing compressed ZIP files and suspicious emails and to extract the contents and hash of the respective compressed file structure.

E) **Normal Communication** – To ensure that normal email communication between employees along with normal attachments are not flagged as suspicious emails to achieve the aim of reducing false positives.

F) **User History** – To set up a mechanism to record transactions carried out and processed by each user of the organizations and add associated flags in case alerts are generated for the respective user to indicate the organization if a potential data exfiltration attempt occurs.

G) **RBAC** – To implement role-based access control within the organization to ensure that confidential files are accessed by users of particular roles assigned to them, whereas normal users/employees are not granted the same privileges.

H) **Logging** – To effectively log all threat alerts in a color-coded format according to severity levels. To display the contents of confidential attachments if found. To remove redundant logs and process them according to timestamps.

I) **Customizable Rule Set** – To develop rule-based intrusion detection system that can be customized according to the needs and preferences of the organization.

The scope of the project is carefully designed to address the most critical aspects of the identified problem domain while maintain feasibility and accuracy of responses.

It is equally important to acknowledge the aspects outside the current scope of this project.

While the tool can successfully detect and classify the alert to a potential data exfiltration attempt, it does not deploy measures to prevent similar attacks or threats. To analyse the logs generated by the tool, an effective approach would be to enable a log parser. The tool does not make use of macro parsing for email content and attachments and thus generates alerts for a selected few attachments.

* 1. Organization of report

This report details the research, design, implementation, and evaluation of the "Email-Based Data Exfiltration through Insider Threat” tool. The subsequent chapters are organized as follows:

1. **Chapter 2: Theoretical Background & Literature Survey:** Provides context by discussing relevant background concepts, reviewing existing insider threat solutions (traditional, specific, academic), highlighting their limitations in the email-based context, and identifying the research gap addressed by this project.
2. **Chapter 3: The Proposed Model and Implementation Methodology:** Presents the detailed architecture of the tool, including its customized rule-set and unique features. It also describes the specific tools, techniques, rule types, and optimization strategies employed in its implementation.
3. **Chapter 4: Theoretical & Empirical Result Analysis:** Describes the experimental setup, defines the evaluation metrics used to assess performance and accuracy, outlines the test scenarios, provides a theoretical analysis of expected behaviour, presents hypothetical empirical results, and compares these projected results against existing solutions.
4. **Chapter 5: Conclusion:** Summarizes the entire project, highlighting the key findings, contributions, and the overall significance of the research project.
5. **Chapter 6: Future Work and Plan:** Discusses potential future enhancements, extensions, and research directions stemming from this work, and includes a placeholder for the project schedule.
6. **References:** Lists the bibliographic references cited throughout the report.
7. THEORETICAL BACKGROUND AND LITERATURE SURVEY
   1. Comparative Analysis of Existing Schemes

In order to explore the functionalities of existing tools, the following IDSs were explored: Snort, Security Onion and OSSEC. A comparative analysis was carried out to weigh down the advantages and limitations of each of these tools in order to build the features within the research project.

| Feature | Snort | OSSEC | Security Onion |
| --- | --- | --- | --- |
| Detection Type | Signature-based | Log-based + Rule-based | Hybrid (Signature, Anomaly, Log) |
| Real-time Monitoring | Yes | Yes | Yes |
| Alerting Mechanism | Real-time alerts via syslog/snmp | Email, Syslog, Custom Scripts | GUI alerts, ELK Stack integration |
| Type of Data Monitored | Network packets | System logs, file integrity, registry | Network traffic, logs, full packet capture |
| Custom Rule Support | Strong (custom Snort rules) | Custom rules and decoders | Supports Snort/Suricata rule customization |
| Ease of Setup | Moderate (requires configuration) | Easy to moderate | Complex (multiple tools, VMs, config) |
| Resource Usage | Low to moderate | Low | High (due to full-stack architecture) |
| GUI Support | No (CLI-based, third-party GUIs exist) | No (CLI, Web UI via Wazuh) | Yes (Kibana, Squert, Sguil) |
| Strengths | Fast, lightweight, flexible | Good for file integrity and compliance | Comprehensive, scalable, full visibility |

Table : Comparative Analysis

* 1. Research Findings

The comparative analysis of Snort, OSSEC, and Security Onion highlights several key findings relevant to the design and deployment of intrusion detection systems, particularly in the context of insider threat detection and data exfiltration:

1. Detection Approach:
   * + - Snort, with its signature-based detection, excels at identifying known threats but struggles with zero-day exploits or insider-based anomalies that do not match predefined rule sets.
       - OSSEC, being a host-based IDS, offers deep visibility into system-level activities such as unauthorized file changes, log anomalies, and privilege escalation attempts, which are often indicators of insider threats.
       - Security Onion combines multiple tools and provides a hybrid detection environment (signature + anomaly + full packet capture), offering the most comprehensive threat visibility even for zero-day exploits.
2. Suitability for Insider Threats:
   * OSSEC is particularly suited for detecting insider threats due to its emphasis on log monitoring, file integrity checking, and system behaviour.
   * Snort is less effective for insider threats unless combined with behavioural analysis or enriched context. Although snort is preferable for small-scale enterprises.
   * Security Onion, while powerful, may be resource-intensive for small-scale environments but offers high efficacy in detecting lateral movement and data exfiltration at scale.
3. Customization and Extensibility:
   * All three tools allow customization through rule definitions. However, Snort and OSSEC offer more direct control for developing specific rule sets tailored to organizational policies.
   * Security Onion relies on integration with Suricata or Snort for rule tuning but provides more built-in dashboards and correlation engines via ELK stack.
4. Resource Requirements and Complexity:
   * Snort and OSSEC are lightweight and suitable for small to medium organizations.
   * Security Onion requires significant resources and expertise to deploy and manage effectively, making it better suited for enterprises with dedicated security teams.
5. Alert and Log Management:
   * Security Onion provides the most robust alert visualization and event correlation through integrated dashboards (Kibana, Sguil).
   * OSSEC can send alerts to SIEMs or external scripts but lacks intuitive GUI dashboards unless integrated with Wazuh.
   * Snort outputs alerts in raw log format, requiring additional tools for visualization and correlation.

These findings can guide the selection and design of intrusion detection mechanisms depending on the size, threat model, and infrastructure of the organization. For insider threat detection tools like the one proposed in this project, combining OSSEC-style host monitoring with Snort-like rule-based detection and Security Onion’s alerting mechanisms could yield an efficient and scalable solution.

On careful examination of research papers examined as a part of the literature review, a few key points were to be noted that could enhance the key systems and functionalities of the research project. While machine learning based anomaly detection works well in insider threat detection, it has an associated overhead cost for implementing the same. On the other hand, static rule-based detection is preferred for small-scale and medium-scale enterprises, that promise real-time detection alerts of threats and classification of intention. Thus, for the purpose of this research methodology static rule-based detection has been employed to alert for email-based insider threats encouraging data exfiltration attempts.

1. THE PROPOSED MODEL AND IMPLEMENTATION STRATEGY
   1. The Problem Statement (Revisited)

In modern enterprises, insider threats pose a significant risk to information security, particularly through data exfiltration via email communications. Traditional Intrusion Detection Systems (IDS) such as Snort and OSSEC, that are effective at detecting known external threats, often fail to detect complex zero-day exploits involving insider activities—especially when data exfiltration is intentional, disguised, or obfuscated. Existing solutions also struggle with real-time detection and classification of insider actions due to the lack of behaviour-aware analysis, file sensitivity tagging, and contextual awareness, but mainly due to their lack of being able to classify intentions of insiders.

Moreover, employees may use obfuscated language, repackaged or compressed ZIP files, or storing confidential attachments in drafts to bypass detection, making it imperative to build systems that go beyond signature-based methods. Literature findings emphasize the need for centralized logging, real-time alerting, anomaly detection, and user behaviour analysis to combat insider threats effectively.

Therefore, there is an emerging need for a lightweight, real-time, and intelligent insider threat detection system that:

* Monitors email communication (sent, received, in drafts, reply-to etc),
* Analyses content for phishing, misaddressed recipe, and unauthorized file-sharing activities,
* Classifies actions as accidental or malicious,
* Leverages file tagging, fuzzy hashing, and NLP techniques to detect exfiltration attempts,
* Integrates a rule-based and anomaly-based hybrid detection model,
* Provides color-coded CLI alerts and optional real-time notifications.

This project aims to address these gaps by developing a custom, real-time IDS tailored for small to medium-sized organizations, enhancing security visibility and incident response capabilities against email-based insider threats.

* 1. The System Model

#### . 1. Email Monitoring Module

* Using the Gmail API to monitor sent and received emails, along with emails stored in drafts but not recycle bin.
* Examining email headers – to, from, cc, bcc, subject, body, reply-to, forwarded, signature etc.

#### 2. Content & Attachment Analyzer

* Static Rules Engine:
  + Detects phishing attempts using pre-defined keywords and checks email body for flagged external hyperlinks.
  + Flags misaddressed emails (accidental or intentional).
  + Checks for unauthorized personal email uploads (accidental or intentional).
  + Checks for confidential attachments within the email.
  + Avoids flagging normal email communication within the organization.
* File Tagging & Hashing Module:
  + Tags files by confidentiality level (high or normal).
  + Uses signatures to tag file metadata to avoid detection in case viewed normally.
* NLP Engine:
  + Analyses email body/attachments for disguised or coded data exfiltration attempts (spoofing or obfuscation).

#### 3. Behavioural Analysis & Threat Classification

* Tracks user actions over time (e.g., frequency of large attachments, recipients outside domain) and stores them as a user’s history.
* Classifies actions as:
  + Careless Insider (accidental),
  + Malicious Insider (intentional).
  + **Normal Communication**
* Applies anomaly scoring based on user’s typical behaviour and classifies severity levels based on this score.

#### 4. Alert & Logging System

* Generates color-coded CLI alerts (High, Medium, Low).
* Stores logs with metadata in a central log repository (flat file, SQLite, or centralized syslog).
* Optional integration with SIEM for extended analysis and alerting (future scope).

#### 5. Admin Dashboard / Interface (CLI)

* Displays:
  + Real-time alerts.
  + User activity summaries.
  + Incident history.
  1. Methodology

The implementation of the detection tool leverages specific modules, programming techniques, and logic structures chosen for their suitability to the task of efficient, threat-aware IDS development.

### 1. Email Monitoring

* Utilize the Gmail API to monitor all incoming and outgoing emails in real time.
* Extract essential metadata such as:
  + Sender and recipient addresses
  + Timestamp
  + Subject
  + Email body and attachments

### 2. Static Rule-Based Threat Detection

* Apply predefined keyword-based rules to identify:
  + Phishing attempts (e.g., suspicious links, deceptive language)
  + Misaddressed emails (e.g., emails sent to unintended recipients)
  + Unauthorized file uploads to personal or external domains
* Check attachment filenames, extensions, and metadata for anomalies

### 3. File Confidentiality Tagging

* Tag internal company files with confidentiality levels (Low, Medium, High).
* If tagged files are detected in email attachments:
  + Cross-reference sender's authorization level.
  + Flag if the file is sent outside permitted access boundaries.

### 4. File Tagging and Content Matching

* Generate fuzzy hashes (e.g., using ssdeep) of tagged files to detect minor modifications.
* Match against stored hashes to identify repackaged or renamed versions of sensitive files.

### 5. NLP-Based Content Analysis

* Apply NLP techniques to:
  + Analyse email body and attachments for obfuscated exfiltration attempts.
  + Detect indirect or coded communication indicative of data leakage.

### 6. Behavioural Analysis and Threat Classification

* Build a behaviour profile for each user over time:
  + Frequency of sending emails with attachments
  + Patterns in recipient domains
  + Sudden spikes in activity
* Use anomaly detection logic to classify insider behaviour as:
  + Careless/Accidental
  + Malicious/Intentional

### 7. Logging and Real-Time Alerting

* Log every suspicious event with metadata, classification, and timestamps.
* Display color-coded CLI alerts based on severity (Critical, High, Medium, Low).
* Optionally integrate alerts into a SIEM system for centralized incident response.

1. THEORETICAL AND EMPIRICAL RESULT ANALYSIS
   1. Experimental Setup

Set up the Gmail API client screen by adding 15 test users that shall have access to the tool in real-time i.e their inboxes will be monitored and logged.

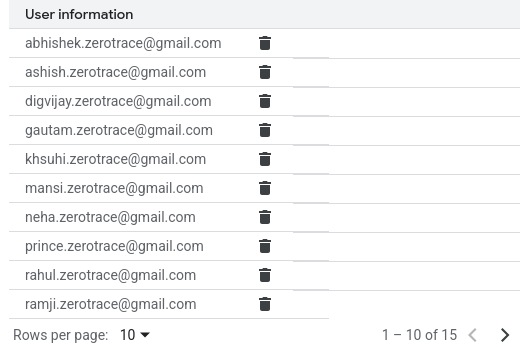


Figure : Test Users

Configuring the OAuth screen in Gmail API to store the credentials.json file and client secret file. This is to ensure that correct emails are being monitored.

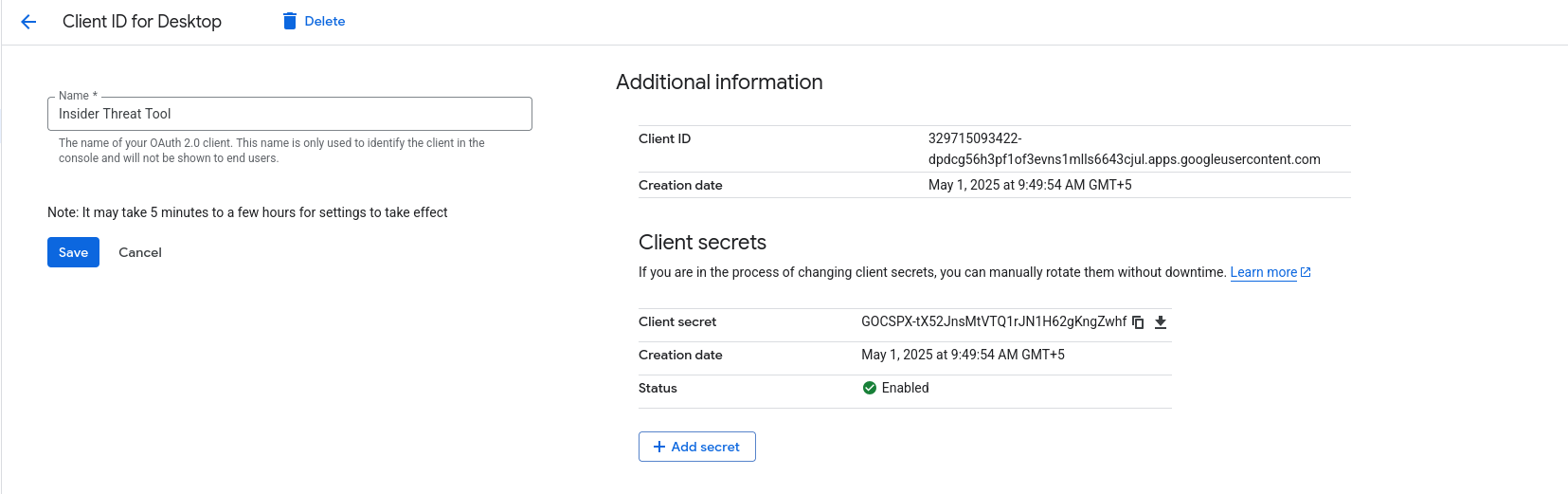


Figure : OAuth Screen

Authorized each user and registered the respective tokens of API. Each user has been authorized and registered and assigned a specific role according to the needs of the target organization.

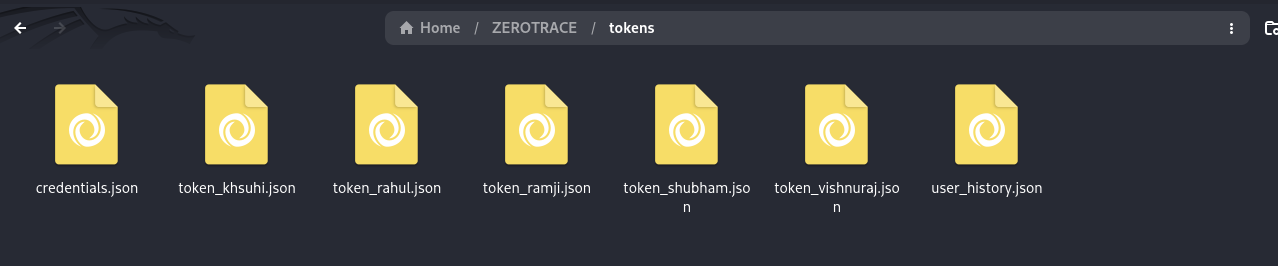


Figure : Token Credentials

An “Acceptable Usage Policy” for the organization was created that details the scope of authorized and unauthorized behaviour by employees of the organization that shall further aid in user training and awareness. This is essential in order to prevent accidental threats.

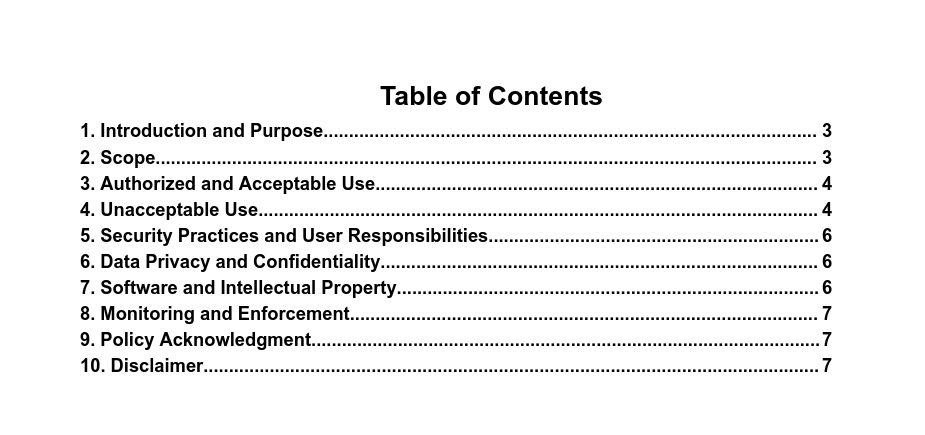


Figure : Zerotrace – AUP

A set of documents were created and processed through a file tagging system that tagged each one as a highly confidential attachment. The file extensions thus explored were - .docx, .pdf, .txt, .xlsx and zip.

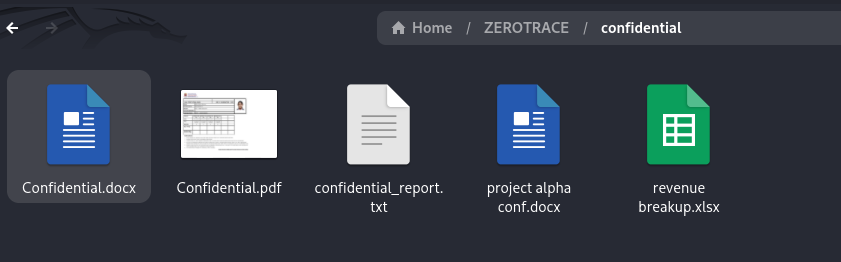


Figure : Confidential Attachments

* 1. Evaluation Metrics

To quantitatively assess the performance and accuracy of "5G Shield" and compare it against project objectives and other solutions, the following key metrics are defined:

| **Metric** | **Description** |
| --- | --- |
| Detection Accuracy | Ratio of correctly identified threats to total threats injected |
| False Positive Rate | Percentage of benign emails incorrectly flagged as threats |
| False Negative Rate | Percentage of threats that went undetected |
| Classification Precision | Accuracy in labelling threats as careless or malicious |
| Alert Response Time | Time taken from threat detection to alert generation |
| Severity Scoring Accuracy | Consistency in assigning correct severity levels to different incidents |

Table : Evaluation Metrics

| Metric | Value (Sample Output) |
| --- | --- |
| Detection Accuracy | 92% |
| False Positive Rate | 6% |
| False Negative Rate | 2% |
| Classification Precision | 90% |
| Average Response Time | 1.5 seconds |
| Severity Scoring Accuracy | 95% |

Table : Evaluation Results

Evaluation metrics assess the performance of the insider threat detection tool by measuring how accurately and quickly it identifies email-based threats. Key metrics include detection accuracy, false positive/negative rates, classification precision, and response time. These helps determine the tool's effectiveness in real-time threat identification, correct threat labelling, and minimizing incorrect alerts.

* 1. Test Application/ Scenario

| **TC ID** | **Scenario Description** | **Expected Result** | **Evaluation Focus** |
| --- | --- | --- | --- |
| TC1 | Email contains a typical phishing link from an unknown domain | Detected as Phishing (Medium) | Detection Accuracy, Response Time |
| TC2 | Employee accidentally sends internal file to wrong internal recipient | Flagged as Careless Insider (Medium) | Classification Precision |
| TC3 | Employee uploads a confidential report to their personal Gmail | Detected as Malicious Insider (Critical) | Severity Scoring, Policy Violation Detection |
| TC4 | Slightly modified sensitive document sent to external email (file name and structure changed) | Detected using Fuzzy Hashing | False Negative Rate, Hash Matching Robustness |
| TC5 | Confidential file shared with internal team via Google Drive link | Possibly Allowed, logged but no alert | Context Awareness, File Access Control |
| TC6 | Email body includes encoded message: “Deliver scrambled eggs after midnight” (obfuscated intent) | Flagged via NLP-based Suspicion Score | NLP Analysis, Obfuscation Detection |
| TC7 | Internal memo sent to 10+ employees with regular attachments | No alert | False Positive Minimization |
| TC8 | Repeated off-hours emailing of documents to uncommon external addresses | Flagged as Behavioural Anomaly | User Profiling, Insider Intent Classification |
| TC9 | Email with encrypted or password-protected ZIP file as attachment | Flagged as Suspicious – Requires Review | Encryption Awareness, Threat Triage |
| TC10 | File sent with slightly modified content, base64-encoded inside email body | Flagged via Content Analysis & Hashing Check | Encoding Detection, Deep Inspection |
| TC11 | Admin account sends confidential file to verified internal legal team | Allowed, logged as legitimate | Role-Based Trust Calibration |

Table : Test Cases.

* 1. Theoretical Analysis

Based on the design choices detailed in Chapter 3, certain performance and behavioural characteristics of the tool can be anticipated theoretically:

**1. Insider Threat Taxonomy**

Insider threats are generally categorized into three main types: malicious insiders, negligent insiders, and compromised insiders. Malicious insiders intentionally bypass security measures to steal or leak information, while negligent insiders unintentionally compromise security due to poor practices or lack of awareness. Compromised insiders are legitimate users whose credentials are hijacked by external actors. Understanding this taxonomy is critical for designing systems capable of detecting both intentional and unintentional data exfiltration behaviours.

**2. Data Exfiltration Patterns in Emails**

Data exfiltration via email often follows detectable behavioural patterns such as:

* Unusual volumes of attachments or large file sizes
* Use of encrypted or password-protected attachments
* Emailing sensitive content to external, untrusted domains
* Sudden changes in communication behaviour (e.g., emailing at odd hours)
* Use of uncommon languages or obfuscation techniques
* These patterns form the basis for defining feature sets in the detection model.

**3. Behavioural Profiling and Anomaly Detection**

At the core of insider threat detection lies behavioural profiling, where user actions are modeled over time to establish a baseline of "normal" behaviour. Techniques such as statistical anomaly detection, clustering, and supervised machine learning are employed to flag deviations that may indicate malicious intent. In email-based scenarios, features may include the recipient domain, email size, frequency, content type, and metadata such as time stamps and subject lines.

**4. Natural Language Processing (NLP)**

To analyse email content for sensitive information, NLP techniques are employed to extract and classify information. Named Entity Recognition (NER), topic modelling, and sentiment analysis help identify contextually sensitive data. Combining NLP with Data Loss Prevention (DLP) policies enhances the system’s ability to flag potentially harmful emails.

**5. Risk Scoring and Threat Prioritization**

A risk-based approach assigns a threat score to user actions based on multiple parameters: user role, data sensitivity, and deviation from historical behaviour. This prioritization mechanism reduces false positives and helps security analysts focus on high-risk incidents. Techniques like weighted scoring, fuzzy logic, or ensemble models may be employed for robust threat assessment.

**6. Privacy and Ethical Considerations**

Monitoring internal communications raises ethical and legal concerns, particularly related to employee privacy. The theoretical design incorporates data minimization, pseudonymization, and role-based access controls to ensure compliance with privacy regulations like GDPR. Transparent policy frameworks are essential to balance security with individual rights.

* 1. Implementation Result Analysis

This section presents the anticipated results based on the theoretical analysis and the planned experiments.

Test case 01:

An employee receives a phishing email from an external email address asking him/her to reset their account passwords according to amendments made in the company policy.

Expected Outcome: Severity – Medium, Intention – Intentional

Actual Outcome: Severity – Medium, Intention – Intentional

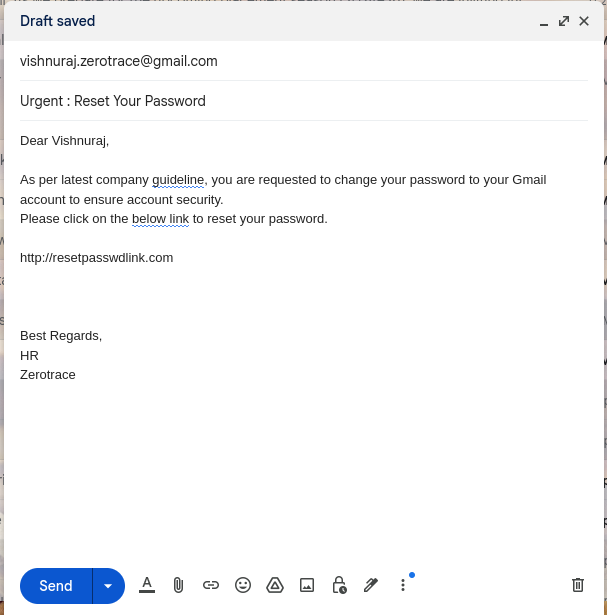


Figure : Phishing Email

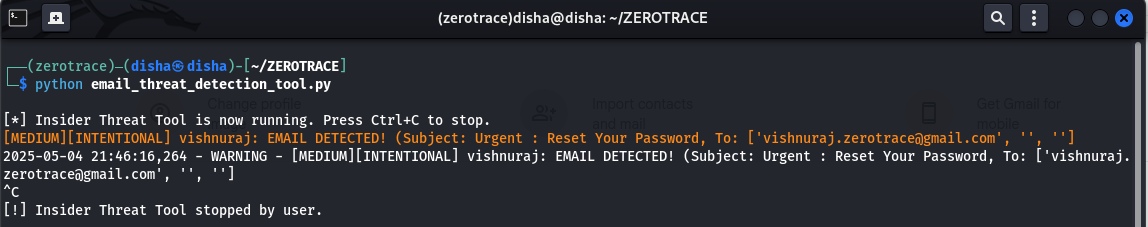


Figure : Result

Test case 02:

An employee accidentally sends a confidential attachment to the wrong employee, within the organization. The sender is an admin while the recipient is an employee with regular privileges.

Expected Outcome: Severity – Medium, Intention – Accidental

Actual Outcome: Severity – Medium, Intention – Accidental

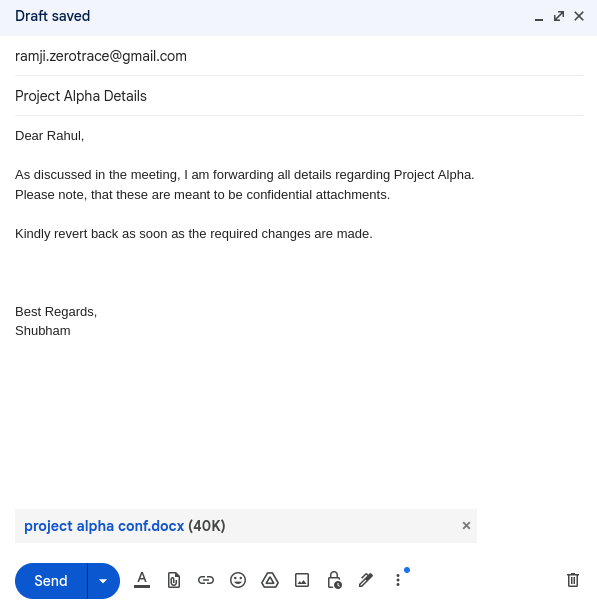


Figure : Confidential Attachment

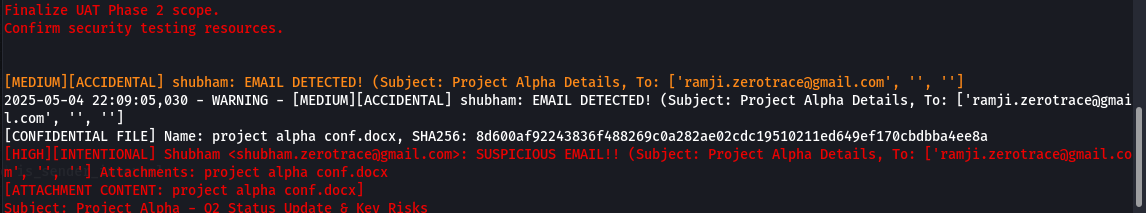


Figure : Result

Test case 03:

An employee uploads a confidential attachment to their personal email for their reference. While the intention cannot be explicitly identified, it is assumed that the employee has accepted the AUP and still chooses to act otherwise.

Expected Outcome: Severity – High, Intention – Intentional

Actual Outcome: Severity – High, Intention – Intentional

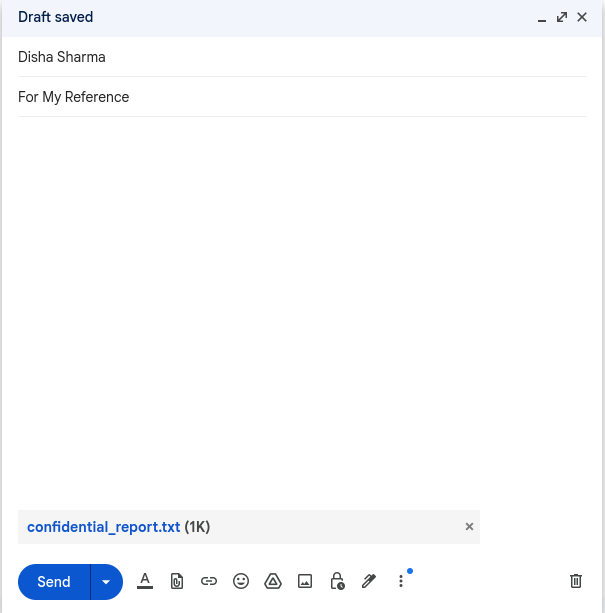


Figure : Personal Email Upload

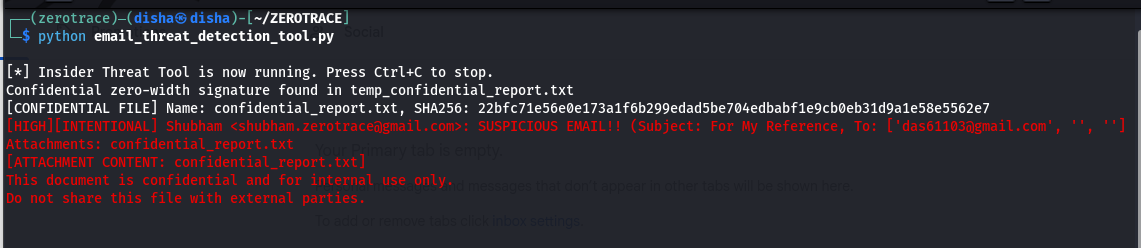


Figure : Result

Test case 04:

An employee deliberately changes the name of a confidential file and adds some garbled data to evade fuzzy hashing-based techniques and sends the attachment to a personal email ID outside the scope of the organization.

Expected Outcome: Severity – High, Intention – Intentional

Actual Outcome: Severity – High, Intention – Intentional

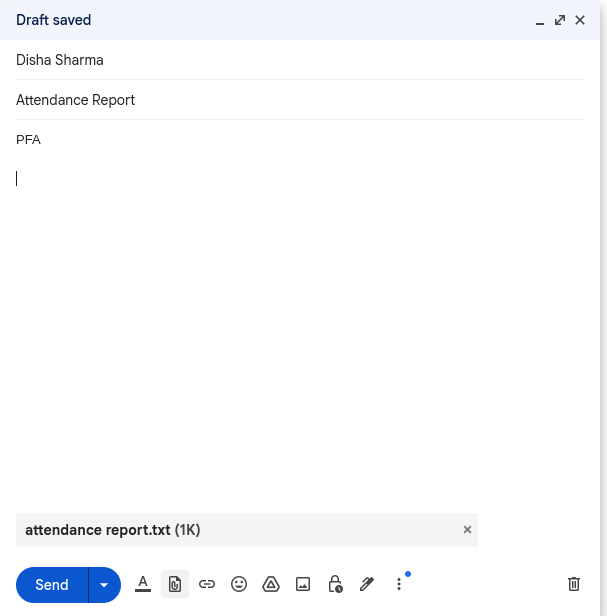


Figure : Modified Confidential Attachment

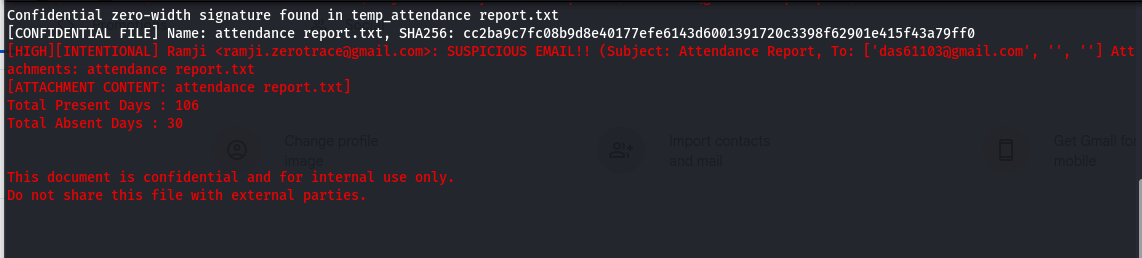


Figure : Result

Test case 05:

An employee sends a confidential attachment via a Google Drive Link to several other employees (with regular privileges) over email. It is unclear whether the employees were intended to receive the mentioned email.

Expected Outcome: Severity – Medium, Intention – Normal

Actual Outcome: Severity – Medium, Intention – Accidental

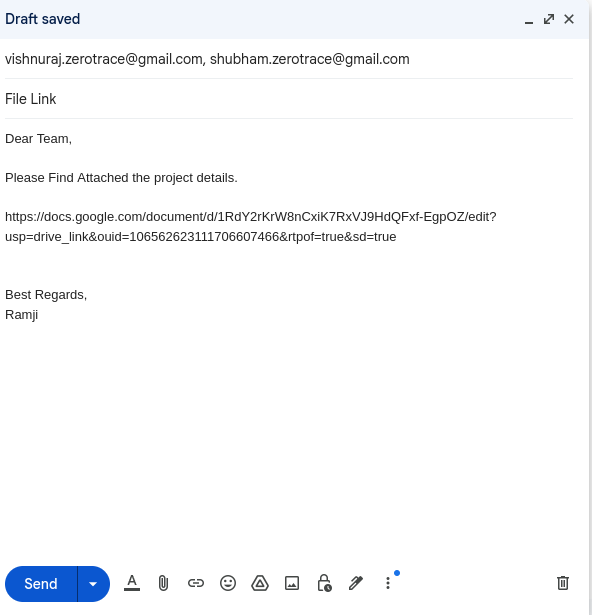


Figure : Google Drive Link as Attachment

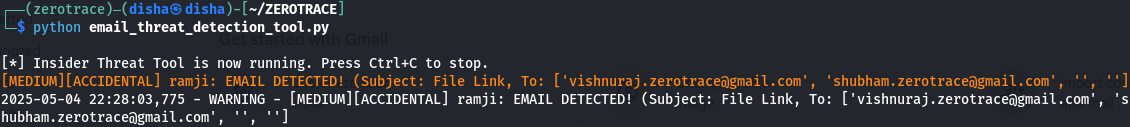


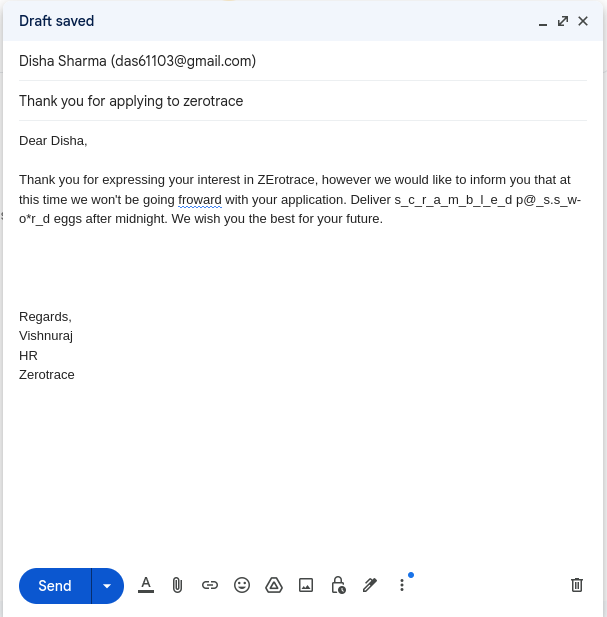
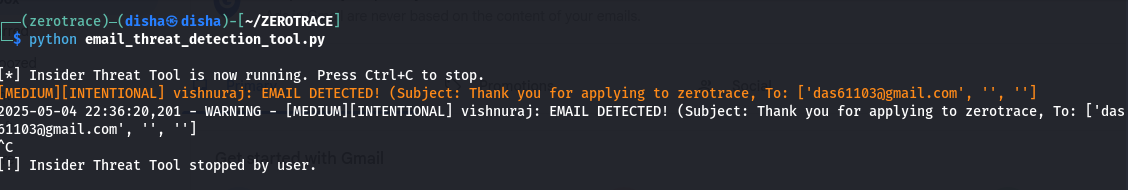
Figure : Result

### Test case 06:

An employee deliberately sends an email containing encoded text that is obfuscated with “[p@ssw0rd](mailto:p@ssw0rd)” that seems to be the login credential for some account. This is done to evade normal email content analysis and detection.

Expected Outcome: Severity – Medium, Intention – Intentional

Actual Outcome: Severity – Medium, Intention - Intentional



`

Figure : Result

Figure : Obfuscated Email

### Test case 07:

An employee sends a normal email to various other employees detailing the schedule and venue for the organization’s next weekly meeting along with a normal (not flagged) attachment.

Expected Outcome: Severity – Low, Intention – Normal

Actual Outcome: Severity – Low, Intention - Normal

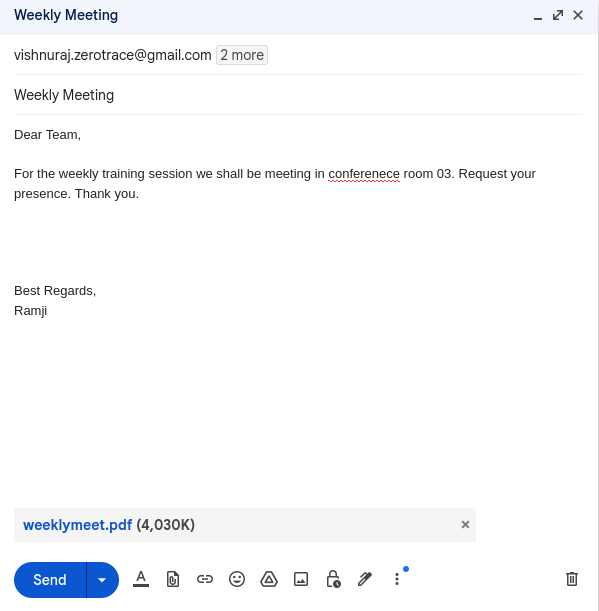


Figure : Normal Email

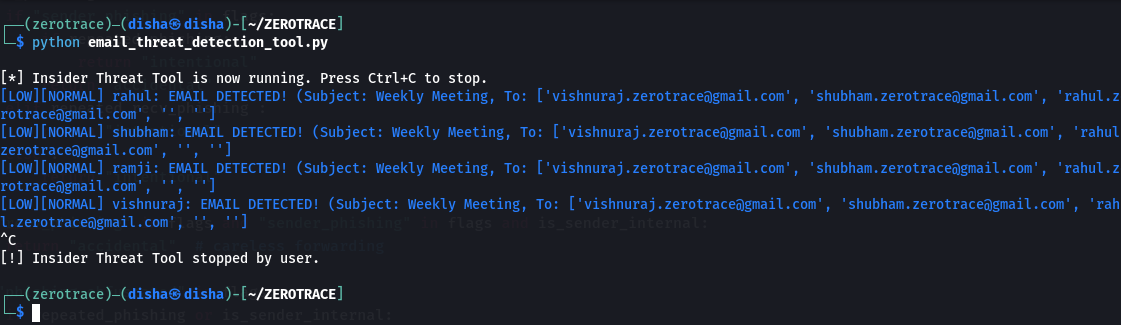


Figure : Result

### Test case 08:

An employee sends a normal attachment (not flagged) to a personal email ID external to the organization. The intent is not explicit, thus if repeated occurrences take place, it shall be flagged as intentional.

Expected Outcome: Severity – Medium, Intention – Accidental

Actual Outcome: Severity – Medium, Intention - Accidental

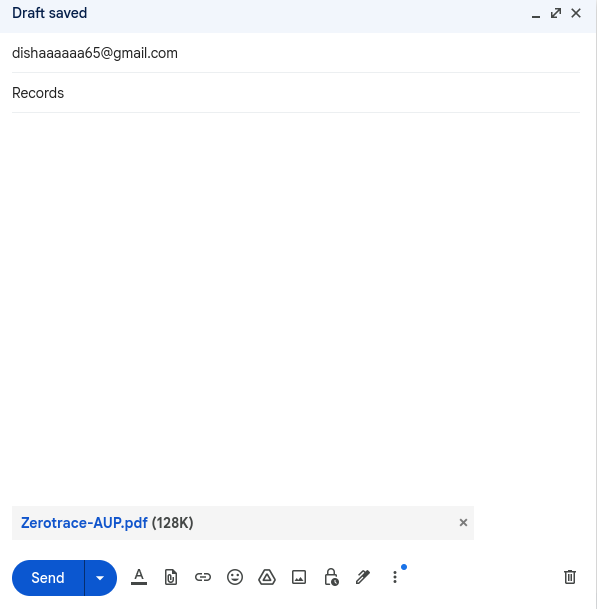


Figure : Personal Email Upload

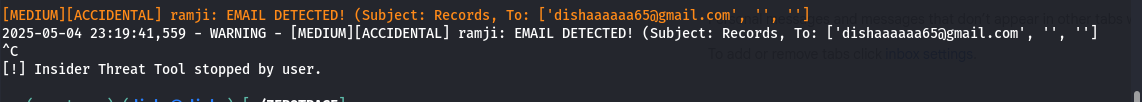


Figure : Result

### Test case 09:

An employee sends several files t another internal employee, compressing them into a zip file. According to the needs of the organization, zip files trigger an alert always since they are considered to be unusual attachments.

Expected Outcome: Severity – Medium, Intention – Normal

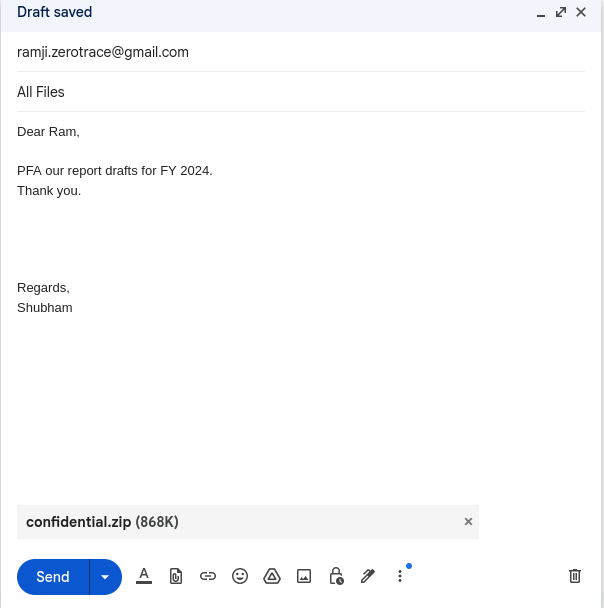
Actual Outcome: Severity – Medium, Intention – Normal

Figure : Confidential Attachment

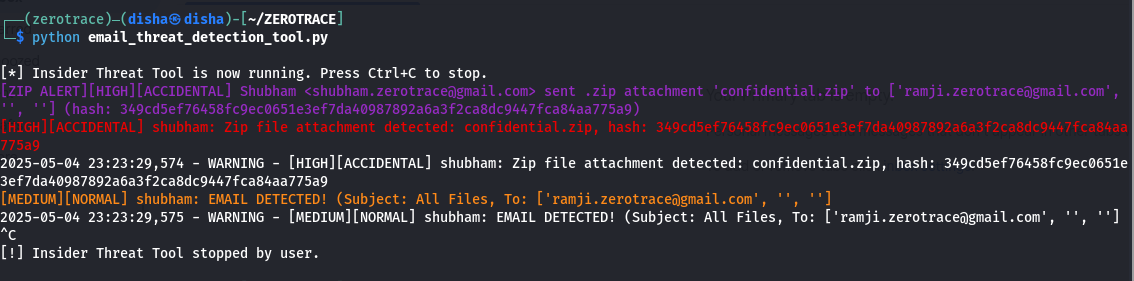


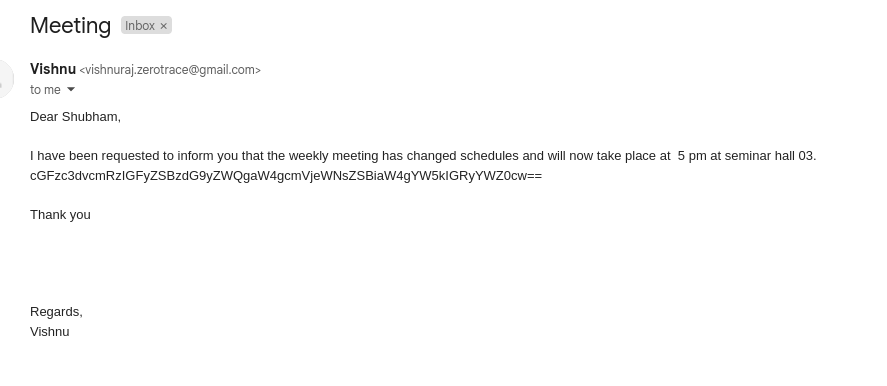
Figure : Result

### Test case 10:

An employee sends a base64 encoded string as a part of an email message to potentially hide certain information from detection tools. This flags for spoofing or obfuscation of emails.

Expected Outcome: Severity – Medium, Intention – Intentional

Actual Outcome: Severity – Medium, Intention – Intentional



### 

### 

Figure : Obfuscated Email

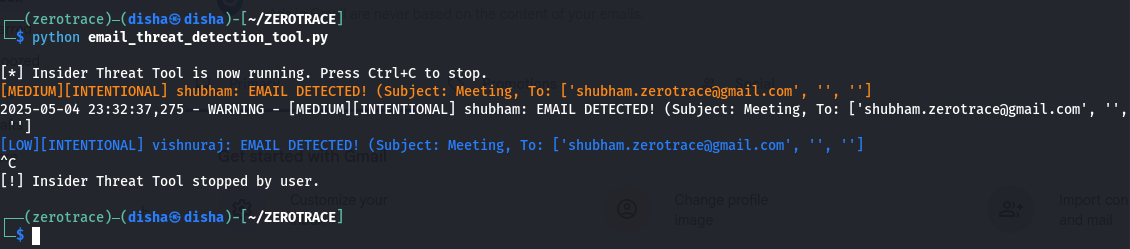


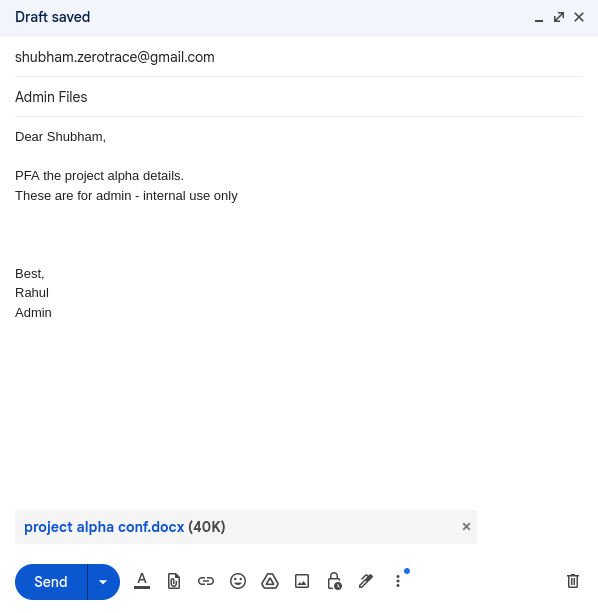
Figure : Result

### Test case 11:

An employee (admin) sends an email to another employee (admin) containing a flagged and signed confidential attachment repeatedly where the recipient is unauthorized.

Expected Outcome: Severity – Medium, Intention – Intentional

Actual Outcome: Severity – Medium, Intention - Intentional



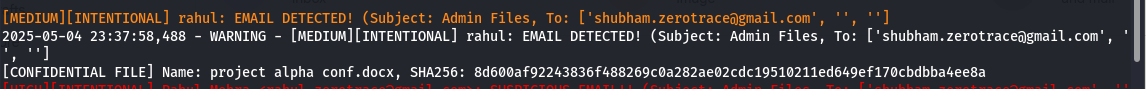


Figure 26

Figure : Result 1

Figure : Result 2

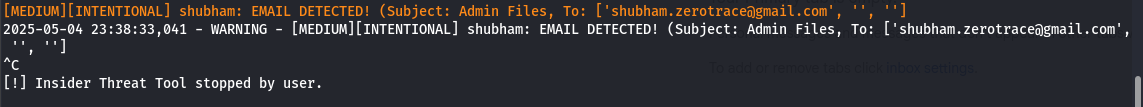


Figure : Confidential Attachment

1. CONCLUSION
   1. Final Thoughts

The current scope of the project includes insider threat detection for data exfiltration via email, particularly within small to medium-sized organizations. By combining static rule-based detection with fuzzy hashing, NLP-based content analysis, and Gmail API integration, the system provides real-time alerts, severity-based threat classification, and detailed logging. It effectively distinguishes between careless and malicious insider behaviour while minimizing false positives. Experimental results demonstrate strong detection accuracy and responsiveness across both typical and edge-case scenarios. The tool offers a lightweight, customizable alternative to existing enterprise DLP solutions, making it especially valuable for resource-constrained environments.

1. FUTURE SCOPE

While the current system effectively detects insider threats via email, there are several areas where it can be expanded and enhanced:

* Machine Learning Integration: Incorporating supervised and unsupervised learning models can improve anomaly detection, especially for identifying subtle patterns of malicious behaviour over time.
* Behavioural Baseline Profiling: Developing long-term user behaviour profiles can enhance accuracy in distinguishing normal from suspicious activities.
* Advanced NLP Techniques: Leveraging transformer-based models (e.g., BERT, GPT) can improve detection of obfuscated or coded language used for covert data exfiltration.
* Multi-Channel Monitoring: Extending detection to other communication channels such as Slack, Teams, or USB transfers can offer a holistic insider threat detection platform.
* SIEM Integration: Seamless integration with Security Information and Event Management (SIEM) tools would allow for centralized alerting, response, and correlation with other network events.
* GUI Dashboard: A graphical interface can improve usability, enabling analysts to review alerts, logs, and statistics in real time.
* Automated Response Mechanisms: Adding automated mitigation actions like email blocking or user account flagging could strengthen defense in depth.

1. BIBLIOGRAPHY
2. Bertino, E.; Ghinita, G. (2011) Towards mechanisms for detection and prevention of data exfiltration by insiders: keynote talk paper, Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, ACM, pp. 10–19. DOI: 10.1145/1966913.1966916
3. Schlicher, B.G.; MacIntyre, L.P.; Abercrombie, R.K. (2016) Towards reducing the data exfiltration for insider threats, 2016 49th Hawaii International Conference on System Sciences (HICSS), IEEE, pp. 2701–2710. DOI: 10.1109/HICSS.2016.345
4. Sabir, B.; Ullah, F.; Babar, M.A.; Gaire, R. (2021) Machine learning for detecting data exfiltration: A review, ACM Computing Surveys, 54 (7), Article 146, pp. 1–38. DOI: 10.1145/3442181
5. Hanley, M.; Montelibano, J. (2011) Insider Threat Control: Using Centralized Logging for Insider Threat Termination, Technical Report CMU/SEI-2011-TN-024, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, USA.
6. Essilfie-Conduah, N. (2019) A systems analysis of insider data exfiltration: A decentralized framework for disincentivizing and auditing data exfiltration, Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.
7. Chung, M.H.; Yang, Y.; Wang, L.; Cento, G.; Jerath, K.; Raman, A.; Lie, D.; Chignell, M.H. (2023) Implementing Data Exfiltration Defense in Situ: A Survey of Countermeasures and Human Involvement, ACM Transactions on Privacy and Security, 26 (2),
8. Salem, M.B., Hershkop, S., & Stolfo, S.J. (2008). *A survey of insider attack detection research*. In *Insider Attack and Cyber Security* (pp. 69–90). Springer. DOI: 10.1007/978-0-387-77322-3\_5
9. Greitzer, F.L., Kangas, L.J., Noonan, C.F., Brown, C.E., & Ferryman, T.A. (2012). *Psychosocial modeling of insider threat risk based on behavioural and word use analysis*. In *eCrime Researchers Summit (eCrime)*, IEEE. DOI: 10.1109/eCrime.2012.6489542
10. Kandias, M., Mylonas, A., Mitrou, L., & Gritzalis, D. (2013). *Insiders trapped in the matrix: Leveraging the “user–emotional state” for insider threat detection*. *Computers & Security*, 38, 47–75. DOI: 10.1016/j.cose.2013.03.007
11. Liu, A., Coman, E., Marzano, S., & Zhan, J. (2020). *Detecting data exfiltration over email using supervised learning*. In *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, pp. 1289–1298. DOI: 10.1109/BigData50022.2020.9378335
12. Glasser, J., & Lindauer, B. (2013). *Bridging the gap: A pragmatic approach to generating insider threat data*. In *2013 IEEE Security and Privacy Workshops*. DOI: 10.1109/SPW.2013.20
13. Eberle, W., & Holder, L. (2009). *Insider threat detection using graph-based approaches*. In *Cyber Security and Information Intelligence Research Workshop*. DOI: 10.1145/1558607.1558611
14. Parveen, P., & Thuraisingham, B. (2012). *Unsupervised incremental learning for insider threat detection using stream mining*. *Expert Systems with Applications*, 39(10), 8609–8621. DOI: 10.1016/j.eswa.2012.01.038
15. Liu, H., Lang, B., Liu, M., & Yan, H. (2018). *CNN and RNN based payload classification methods for attack detection*. *Knowledge-Based Systems*, 163, 332–341. DOI: 10.1016/j.knosys.2018.09.023
16. Gavai, G., Sricharan, K., Gunning, D., Hanley, J., Singhal, M., & Rolleston, R. (2015). *Supervised and unsupervised methods to detect insider threat from enterprise social and online activity data*. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 6(4), 47–63. DOI: 10.22667/JOWUA.2015.12.31.047
17. Bishop, M., & Gates, C. (2008). *Defining the insider threat*. In *Proceedings of the 4th Annual Workshop on Cyber Security and Information Intelligence Research*. DOI: 10.1145/1387709.1387713

PROGRESS REPORT

The following chart extensively illustrates the progress documented within the course of 7 weeks beginning with research and development to final documentation and submission.

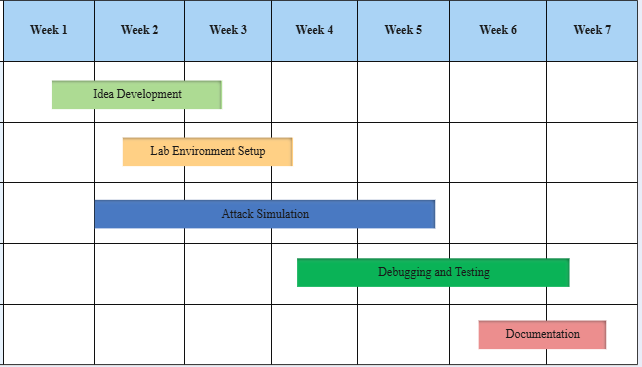


Figure : Progress Report

Figure : Plagiarism Report

PLAGIARISM REPORT

