DATA LEAKAGE DETECTION SYSTEM

A PROJECT REPORT

Submitted by

Yash Saxena 21BCS7073

Sahil Arora 21BCS7498

Nandini 21BCS7633

Hitaansh Maheshwary 21BCS7465

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Chandigarh University

MAY 2025

BONAFIDE CERTIFICATE

Certified that this project report **“DATA LEAKAGE DETECTION SYSTEM”** is the bonafide work of “**YASH SAXENA, SAHIL ARORA, NANDINI, HITAANSH MAHESHWARY”** who carried out the project work under my/our supervision.

SIGNATURE
Dr. Puneet Kumar
HEAD OF THE DEPARTMENT
CSE

SIGNATURE
Banisha Sharma
SUPERVISOR
Assistant Professor CSE

Submitted for the project viva-voce examination held on 20-03-2023

INTERNAL EXAMINER **EXTERNAL EXAMINER**

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Identification of Client / Contemporary issue

Client: Big Companies whose most of the work is relied on the data exchange between different departments. Hospitals who share sensitive data of the patients with the research laboratories. Any organization that shares confidential data with other of their trusted agents and if leaked can result in huge loses to the organization.

Need: Today every organization works on a lot of data which is shared between different departments of the organization or so called the trusted agents and also some of the third-party agents. Thus, the value of an organization is measured using the data it generates and how it uses that data to improve themselves or their services. If this confidential data is leaked then it will result in the deprecation of the value of the organization and its integrity. Therefore, we need to ensure the security of this data.

1.2 Identification of Problem

A number of issues can develop in a data leakage detection system that reduce its effectiveness. False positives and false negatives—where acceptable actions are mistakenly reported as cases of data leakage or real occurrences are overlooked—are frequent problems. Inadequate data classification can make it difficult for the system to recognize and safeguard sensitive information. Data breaches may go undiscovered due to low visibility across several channels, such as email or file-sharing websites. It may be difficult to distinguish between typical and suspect activity if user behavior profiling is inaccurate. Data leakage may become more likely if privileged users who have access to sensitive information are not sufficiently monitored.

The system's capacity to correlate events and identify potential leaks is constrained by a lack of integration with other security technologies. Inadequate employee data security training and awareness might leave workers vulnerable to social engineering attacks or unintended data releases. To find new or undiscovered dangers that might elude conventional detection methods, proactive threat hunting is crucial. Data leakage issues may be handled slowly or incorrectly due to inadequate incident response skills and a lack of well-defined response protocols. Last but not least, skipping routine system upgrades and maintenance makes the system more susceptible to fresh attack vectors and less effective at spotting and stopping data leaks.

1.3 Identification of Tasks

Identify: The identification phase involves determining the requirements for the Data Leakage Detection System. This includes understanding the needs of the people, the technology stack to be used, and the overall scope of the project. Some of the tasks related to this phase are listed below:

- Understanding the different data leakage sources and causes.
- Defining the functional and non-functional requirements for Data Leakage Detection System.
- Defining the user stories and use cases to identify the features required.
- Selecting the appropriate technology stack.

Build: The building phase involves developing the software based on the requirements identified in the previous phase. This involves the following tasks:

- Building a watermarking system that is more efficient than the pre-existing systems and embedding codes in it which can be later used to identify the data.
- Conducting and evaluating the software for a given use case.

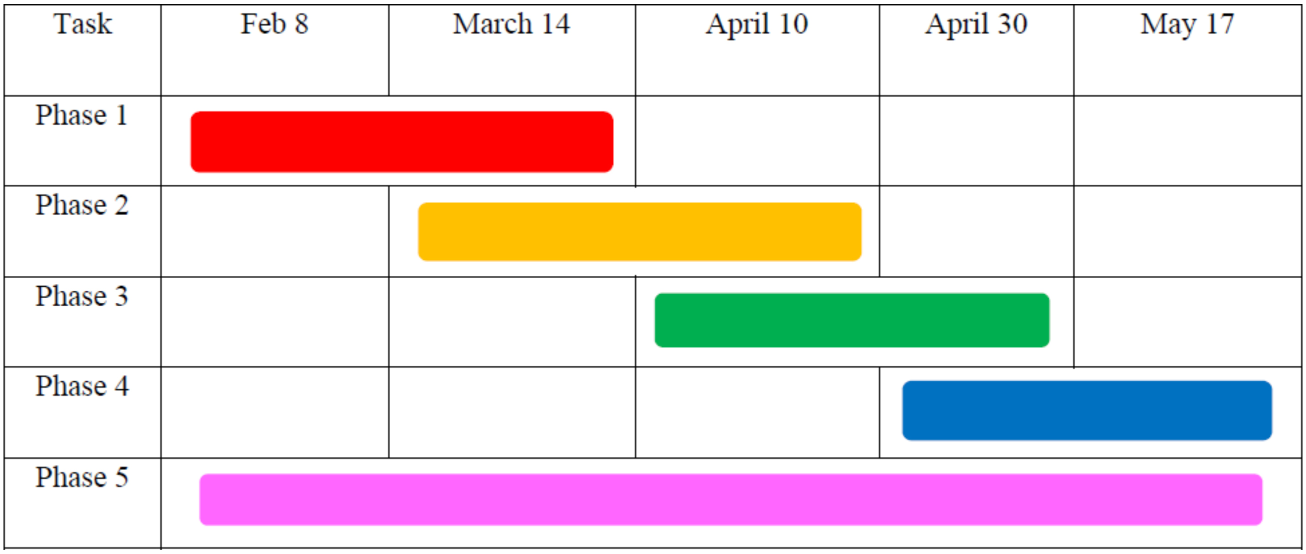
Test: The testing phase involves verifying that the software meets the requirements and functions as intended. This includes the following tasks:

- Conducting unit testing, integration testing, and system testing to ensure that the detection system functions correctly.
- Conducting user acceptance testing to ensure that the software meets the needs of the people outside.
- Conducting security testing to identify and fix any vulnerabilities.
- Conducting performance testing to ensure that the software can handle different use cases. Fixing any bugs and issues identified during testing.

In summary, the identification phase involves determining the requirements for the people outside, the building phase involves developing the program based on these requirements, and the testing phase involves verifying that the program would meet the requirements and functions as intended.

1.4 Timeline

Review of Data Leakage Detection System Calendar
Phase 1- Project Scope, Planning and Task definition [Feb 8 – March 14]
Phase 2- Literature Review [March 14 – April 10]
Phase 3- Preliminary design [April 10 – April 30]
Phase 4- Detailed System Design/Technical Details [April 30 – May 17]
Phase 5- Work Ethics [Feb 8 – May 17]



Task	Feb 8	March 14	April 10	April 30	May 17
Phase 1	🟥	🟥			
Phase 2		🟧	🟧		
Phase 3			🟩	🟩	
Phase 4				🟦	🟦
Phase 5	🟪	🟪	🟪	🟪	🟪

Fig 1.1- Gantt Chart used to illustrate plan of project timeline

1.5 Organization of the Report

Chapter 1: The introduction gives a brief outlook on what the research is about, what it includes, and the problems highlighted whose solutions in turn are fulfilled by the site worked upon. A little research about the topic is done which brings about the surveys that already have occurred, helping to determine and improve our work. Categorizing the tasks that need to be achieved is also part of the introduction which helps us to distinguish and build a framework. Defining timelines and giving a structure of what the report reflects brings about a detailed overview of what the project will show.

Chapter 2: Literature review and background study, an important aspect that reflects the time invested and research done on our project. All the problems previously defined are given a solution that also contributes to the details of the research paper. Some problems may be hard to tackle or there can be issues occurring which reflect the working of the project, adding to the drawbacks. Giving a summary of all that is researched and concluded, along with a clear definition of what is about to be going on with the site. Defining goals or objectives here can give a clear view of what needs to be achieved.

Chapter 3: Design Flow and Process deals with the working of the website. The previous chapter helps us to determine problems, solutions, drawbacks, and every aspect the research concludes giving a clearer objective on what needs to be achieved. The how objective needs to be achieved is dealt with in the design flow. The basic layout, structure, and design of the website are worked upon along with all the coding, development, and features that need to be the part of our product. Alternate solutions are found and the best approach is finalized and worked on to achieve a successful model which clearly fulfills almost every objective.

Chapter 4: Result analysis is the part where the final design along with a proper flowchart is presented and then finally the successfully running product is displayed along with the proper details of management of both the product and report.

Chapter 5: Conclusion and Future work will be the last phase of the report which sums up everything that'll be done till the very end where our working product is achieved along with all the solutions that were once a problem in the beginning. Any changes from original objective or any modified versions of what finally as a product is gained are mentioned in conclusion. Future work includes what else could be done for improvement or betterment, also how can one overcome the drawbacks that are being faced.

These 5 chapters of the report summarizes our whole project giving a third person a detailed yet specific version of the project. Problems, Solutions and every little important detail is talked about helping anyone to understand and go through about how and what is being accomplished.

CHAPTER 2
LITERATURE REVIEW/BACKGROUND STUDY

2.1 Timeline of the Reported Problem

Data leakage is a significant problem in the digital era. The problem is prevalent across various sectors, including healthcare, finance, and government agencies. Data leakage can lead to privacy breaches, financial losses, and reputational damage to organizations. The problem of data leakage has been on the rise since the early 2000s, with the increasing use of digital technologies and the internet. The rise of cloud computing and big data has further exacerbated the problem.

2003: The first major data breach occurs when hackers gain access to customer data at TJX Companies, exposing 45.7 million credit and debit card numbers.

2005: A laptop containing personal information on 28,000 NASA employees is stolen, highlighting the risk of data leakage through portable devices.

2007: The credit reporting agency, Equifax, suffers a data breach that compromises the personal information of 143 million consumers.

2013: Edward Snowden leaks classified information from the National Security Agency, prompting concerns about insider threats and the need for better data protection.

2017: The WannaCry ransomware attack infects hundreds of thousands of computers worldwide, highlighting the importance of proactive measures for detecting and preventing data leakage.

2018: The European Union's General Data Protection Regulation (GDPR) goes into effect, imposing strict requirements for data protection and breach notification.

2020: The COVID-19 pandemic leads to a surge in remote work, creating new challenges for data leakage prevention as employees access sensitive information from outside the office.

This timeline shows that data leakage has been a significant problem for many years, with high-profile breaches occurring across a range of industries and settings. As technology continues to evolve and new threats emerge, it is important to stay vigilant and take proactive measures to protect sensitive data.

2.2 Existing Solutions

Various techniques have been proposed to address the problem of data leakage. These include encryption, access control, data masking, and watermarking. Access control restricts access to data based on user roles and privileges. Data masking involves obscuring sensitive data to prevent unauthorized access. Watermarking involves embedding a unique identifier in data to trace its origin and prevent unauthorized use.

Anomaly-based detection: This approach uses statistical models to detect deviations from normal patterns of behavior, such as unusual file access or network traffic. Anomaly detection can be effective for detecting unknown or previously unseen threats, but may also produce false positives if normal behavior patterns are not well understood.

Content-based detection: This approach uses pattern matching and keyword searches to identify sensitive data, such as credit card numbers, social security numbers, or intellectual property. Content-baseddetection can be highly accurate but may also require manual configuration and maintenance to keep up with evolving data types and formats.

Data loss prevention (DLP) systems: These systems provide a comprehensive approach to data leakage prevention, combining both anomaly and content-based detection techniques with policy-based controls for data handling and encryption. DLP systems can be highly effective but can also be complex and expensive to implement and maintain.

Machine learning-based detection: This approach uses supervised or unsupervised machine learning algorithms to detect patterns of behavior or content that may indicate data leakage. Machine learning-based detection can be highly effective for detecting unknown or emerging threats, but may also require large amounts of data and computing resources to train and maintain the models.

User behavior analytics (UBA): This approach uses machine learning algorithms to analyze user behavior patterns and identify anomalies or suspicious activities, such as accessing sensitive data outside of normal business hours or from an unfamiliar device. UBA can be effective for detecting insider threats or compromised accounts, but may also require careful tuning and monitoring to avoid false positives.

These existing solutions provide a range of options for detecting and preventing data leakage, each with its own strengths and weaknesses. The choice of solution will depend on factors such as the organization's size, resources, and risk tolerance, as well as the types of data being protected and the regulatory requirements that apply.

2.3 Bibliometric Analysis

Recent research has focused on developing machine learning-based approaches for data leakage detection. These approaches involve analyzing data traffic patterns to identify anomalous behavior that may indicate data leakage. However, existing projects have limitations such as high false positive rates and limited scalability. Let us look at some of the projects:

Project proposed an anomaly-based approach for detecting data leakage by analyzing network traffic. The approach achieved high accuracy but had a high false positive rate.

"ML based data leakage detection" by Li et al. (2018) - This project proposed a machine learning-based approach for detecting data leakage in cloud computing environments. The approach achieved high accuracy but had limited scalability.

"Deep learning-based data leakage detection in mobile environments" by Zhang et al. (2020) - This project proposed a deep learning-based approach for detecting data leakage in mobile environments. The approach achieved high accuracy but had limited applicability to other environments.

Some key features of existing data leakage detection systems identified through bibliometric analysis include:

An increasing trend in research activity over time, with a particularly sharp rise in publications since the early 2010s.

A focus on content-based detection techniques, particularly those based on machine learning algorithms and natural language processing.

A growing interest in user behavior analytics and the use of context-based information to improve detection accuracy.

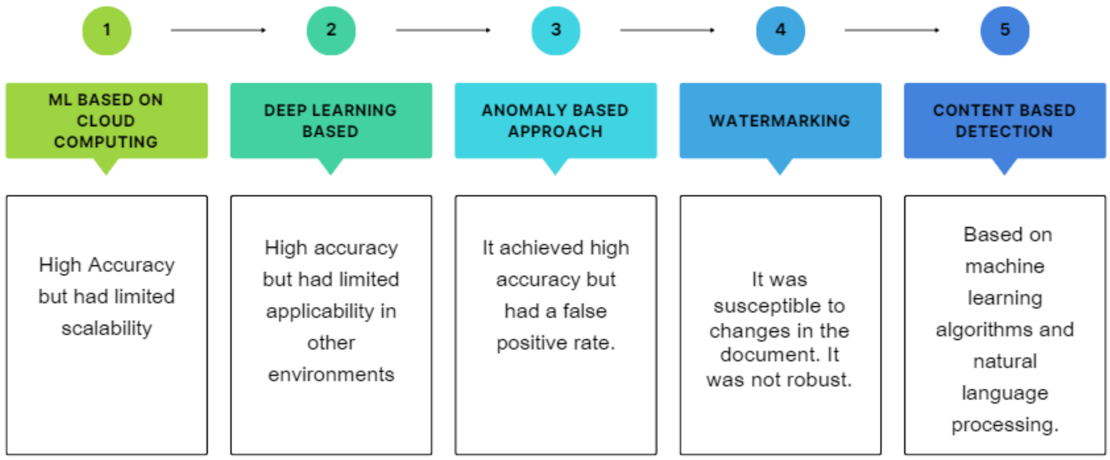
A relatively small number of high-impact publications, indicating a need for more collaborative and interdisciplinary research in the field.

A lack of standardization in evaluation metrics and datasets, making it difficult to compare and replicate results across studies.

A need for more research on the effectiveness of different detection techniques under varying conditions and threat models.

Overall, the bibliometric analysis suggests that while significant progress has been made in developing data leakage detection systems, there are still many challenges to be addressed. Future research should focus on developing more standardized evaluation metrics and datasets, as well as exploring new detection techniques that can address emerging threats and evolving data types.

BIBLIOMETRIC ANALYSIS



1 ML BASED ON CLOUD COMPUTING

High Accuracy but had limited scalability

2 DEEP LEARNING BASED

High accuracy but had limited applicability in other environments

3 ANOMALY BASED APPROACH

It achieved high accuracy but had a false positive rate.

4 WATERMARKING

It was susceptible to changes in the document. It was not robust.

5 CONTENT BASED DETECTION

Based on machine learning algorithms and natural language processing.

2.4 Review Summary

Data leakage is a growing concern for organizations of all sizes and sectors, as the volume and complexity of digital data continue to increase. This literature review has examined the reported problem of data leakage detection and the existing solutions to address this problem. The timeline of the problem indicates a significant rise in research activity since the early 2010s, with a focus on content-based detection techniques such as anomaly detection and machine learning algorithms.

Bibliometric analysis has identified a need for more standardized evaluation metrics and datasets, as well as more collaborative and interdisciplinary research to address the complex and evolving nature of data leakage threats. The existing solutions identified in this review include anomaly-based detection, content-based detection, data loss prevention systems, machine learning-based detection, and user behavior analytics. Each approach has its strengths and limitations, and the choice of solution will depend on the organization's specific needs and risk profile. The goals and objectives of a data leakage detection system should be aligned with the organization's overall security strategy, taking into account regulatory requirements and the need for ongoing monitoring and evaluation. Overall, this literature review highlights the importance of continuous innovation and collaboration in the development of effective data leakage detection systems.

2.5 Problem Definition

Data leakage refers to the unauthorized transfer of sensitive data from within an organization to an external destination. With the increasing volume and complexity of digital data, data leakage has become a significant and growing concern for organizations of all sizes and sectors. Data leakage can lead to financial losses, reputational damage, and legal and regulatory penalties. Detecting data leakage in real-time is crucial to minimizing its impact and preventing further loss of sensitive information. However, existing solutions face challenges in terms of accuracy, scalability, and adaptability to evolving threats.

The problem of data leakage detection is complex and multifaceted, requiring a holistic approach that considers technical, organizational, and human factors. Effective data leakage detection systems must balance the need for detection accuracy with the need for user privacy and usability, while also taking into account regulatory requirements and organizational policies. Addressing these challenges requires ongoing innovation, collaboration, and interdisciplinary research.

2.6 Goals and Objectives

The goal of this project is to develop a machine learning-based approach for data leakage detection that addresses the limitations of existing solutions. The objectives of the project include:

- Analysing data traffic patterns to identify anomalous behaviour that may indicate data leakage

- Improving the accuracy of data leakage detection while reducing false positive rates

- Developing a scalable approach that can be applied in different environments
- Evaluating the performance of the proposed approach using real-world datasets

CHAPTER 3
DESIGN FLOW/PROCESS

3.1 Evaluation and Selection of Specifications/Features

Evaluation and selection of specifications/features involves identifying the critical requirements for the data leakage detection system. This includes determining the types of data to be monitored, the methods of data transmission and storage, the potential attack vectors, and the type of analysis required. The evaluation process requires an in-depth understanding of the system's functional and non-functional requirements, as well as the potential security threats that may affect the system's performance.

To evaluate the specifications/features of the system, it is important to identify the key factors that determine the system's effectiveness, such as accuracy, speed, scalability, and ease of use. These factors can be evaluated by conducting experiments, simulating different scenarios, and testing the system under various conditions.

After evaluating the system's specifications/features, the next step is to select the most suitable ones for the data leakage detection system. This involves analyzing the trade-offs between different features and selecting the ones that best meet the system's requirements. The selected specifications/features should be able to provide a high degree of accuracy, fast detection, scalability, and user-friendliness.

Overall, the evaluation and selection of specifications/features is a crucial stage in the design of a data leakage detection system. It ensures that the system is designed to meet the requirements of the stakeholders and is capable of detecting potential threats effectively.

3.2 Design Constraints

Design constraints are the limitations and restrictions that affect the design process and outcomes of a project. In the context of a data leakage detection system, some design constraints may include:
Resource Limitations: The system must be designed to work within the available resources such as CPU, memory, storage, and bandwidth.

Time Constraints: The system must be designed within the allocated time frame for the project.
Compatibility Constraints: The system must be designed to work with existing technologies and systems that it is intended to integrate with.

Security Constraints: The system must be designed to prevent unauthorized access and protect the privacy of sensitive data.

Scalability Constraints: The system must be designed to handle a large amount of data and be scalable to accommodate future growth.

Cost Constraints: The system must be designed to meet the budgetary constraints of the project.
Legal and Regulatory Constraints: The system must be designed to comply with legal and regulatory

requirements such as data protection laws, privacy regulations, and industry standards.

Design constraints play an essential role in the design process as they help guide the decision-making process and ensure that the system meets the necessary requirements. It is important to identify and prioritize design constraints early in the design process to ensure that they are taken into account during the design, development, and implementation stages. Failure to consider design constraints can result in costly redesigns and delays in the project timeline.

3.3 Analysis of Features and finalization subject to constraints

After identifying the design constraints, the next step in the design process is to analyze the available features and finalize the appropriate ones based on the identified constraints. For the data leakage detection project, the following features will be analyzed and evaluated:

Data Access Control: Access to sensitive data should be restricted to authorized personnel only. The system should be able to enforce access control policies to ensure that data is accessed only by authorized personnel.

Data Encryption: The system should be able to encrypt sensitive data to protect it from unauthorized access. The encryption algorithm used should be strong enough to prevent data breaches.

Data Leak Prevention: The system should be able to detect and prevent data leaks. It should be able to monitor data transfers and detect any unauthorized attempts to transfer sensitive data.

Anomaly Detection: The system should be able to detect anomalies in data transfer patterns. It should be able to detect deviations from normal behavior and raise alerts.

Logging and Auditing: The system should maintain logs of all data transfers and access to sensitive data. The logs should be audited regularly to detect any suspicious activity.

After analyzing the above features, the appropriate ones will be selected based on the design constraints identified earlier. For instance, if the system needs to be implemented in a resource-constrained environment, features such as data encryption may be prioritized over features such as anomaly detection due to their lower computational overhead.

Once the appropriate features have been selected, the next step is to finalize their design and implementation. This involves determining the specific algorithms and techniques to be used to implement the selected features, as well as the integration of these features into the overall system design.

Overall, the analysis of features and finalization subject to constraints is a critical step in the design process for the data leakage detection project, as it ensures that the selected features are appropriate for the identified constraints and will effectively mitigate the risk of data breaches.

3.4 Design Flow

Design flow is the process of defining the steps required to develop a solution to a particular problem. In the case of the data leakage detection system project report, the design flow would include the following steps:

Requirement analysis: In this step, the requirements for the data leakage detection system are gathered and analyzed. This includes understanding the data sources, types of data that need to be protected, and the potential threats to the data.

System design: Based on the requirements, the system is designed to meet the needs of the user. This includes determining the architecture, selecting the algorithms, and deciding on the system components.
Implementation: The system is implemented using the chosen programming language and tools. This step includes writing the code, creating the database, and configuring the system.

Testing: In this step, the system is tested to ensure that it meets the requirements and is functioning as intended. This includes unit testing, integration testing, and system testing.

Deployment: Once the system has been tested and approved, it is deployed in the production environment.
Maintenance: The data leakage detection system must be maintained and updated regularly to keep up with the latest threats and to ensure that it continues to meet the user's needs.

The design flow for the data leakage detection system project report should follow a structured approach that includes each of the above steps. Each step should be well-documented to ensure that the project can be easily replicated and maintained. It is also important to have a clear understanding of the user's needs and requirements at each stage of the design flow to ensure that the final product meets their expectations.

3.5 Design Selection

Design selection is an important aspect of any project, and it is crucial for the success of the project to choose the best possible design. In the case of a data leakage detection system, design selection involves selecting the most appropriate algorithm and approach for detecting data leaks.

There are various design options available for data leakage detection, including rule-based systems, signature-based systems, anomaly detection, and machine learning-based approaches. Each of these approaches has its strengths and weaknesses, and the selection of the most appropriate design depends on the requirements and constraints of the project.

The design selection process involves evaluating the different design options based on various criteria, such as accuracy, efficiency, scalability, and adaptability. The design should be able to detect different types of data leaks, including those that involve structured and unstructured data.

Once the design options have been evaluated, the best design option is selected based on the project requirements and constraints. The selected design should be implemented and tested to ensure that it meets the project's objectives and requirements.

In summary, the design selection process for a data leakage detection system involves evaluating different design options based on various criteria and selecting the most appropriate design based on the project's requirements and constraints.

3.6 Implementation plan/Methodology

Once the design flow and design selection are finalized, the implementation plan and methodology can be formulated. This involves defining the steps to be taken to convert the design into a functional data leakage detection system.

The first step is to set up the required hardware and software infrastructure for the system. This includes identifying the hardware components, such as servers and network devices, and installing the required software, such as the operating system, database management system, and data leakage detection software. Next, the system needs to be configured to meet the design specifications and constraints. This involves setting up the various parameters, thresholds, and rules required for the detection of data leakage. The system should also be tested for its efficiency and effectiveness in detecting different types of data leakage scenarios.

Once the system is configured and tested, it needs to be integrated with the existing IT infrastructure of the organization. This involves identifying the various endpoints and network devices where the data leakage detection software needs to be installed and configuring them accordingly. The system should also be tested for compatibility and stability in the new environment.

Finally, the implementation plan should include a training program for the users and IT staff. This involves educating them on the various features and capabilities of the system and training them on how to use it effectively. The training should also cover best practices for data protection and data leakage prevention. Overall, the implementation plan and methodology should ensure that the data leakage detection system is implemented smoothly and effectively, and that it meets the design specifications and constraints while providing maximum protection against data leakage.

3.7 Flow Diagrams

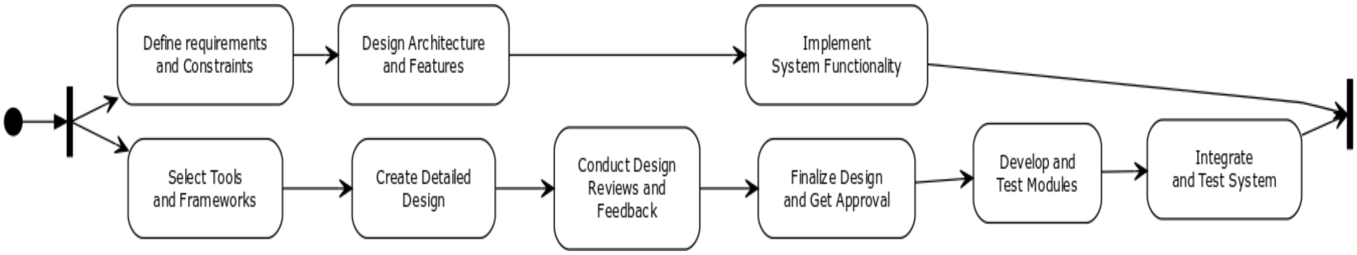

Define requirements and Constraints
Design Architecture and Features
Implement System Functionality
Select Tools and Frameworks
Create Detailed Design
Conduct Design Reviews and Feedback
Finalize Design and Get Approval
Develop and Test Modules
Integrate and Test System

Fig 3.1- Flow diagram for implementation of project

CHAPTER 4
RESULTS ANALYSIS AND VALIDATION

4.1 Implementation of solution

A combination of technical measures, rules, and processes are used to implement solutions in a data leakage detection system in order to detect and stop data leaks. Here are some crucial actions to take into account while putting solutions in place for a data leakage detection system:

- Define a policy for data leakage, create a thorough data leaking policy first, outlining what sensitive information is, how to handle it, and the repercussions of data leakage. All staff members and stakeholders should be informed about this policy.

- Data classification, arrange your information according to its importance and sensitivity. To help focus protection efforts, classify data into different categories (such as internal, public, confidential, and very confidential).

- Strong access controls should be implemented to guarantee that only authorized individuals can access sensitive data. Use role-based access controls (RBAC) and the least privilege principle to impose access restrictions on sensitive data.

- Encrypt sensitive data both at rest and while it is being transmitted. Use robust encryption techniques, and watch out for how encryption keys are handled.

- Tools for data loss prevention (DLP) use data loss prevention technologies to track and stop data leaks. These solutions can recognize and prevent sensitive information from leaving the company via a variety of channels, including email, file transfers, and web uploads.

- Implement user activity monitoring systems to track, examine, and identify user behavior as well as any unusual or unauthorized behaviors. To spot potential data leakage situations, keep an eye on file accesses, data transfers, and user behaviors.

- Monitoring of network traffic, use tools for network monitoring to examine network traffic patterns and spot any irregularities that might point to data leaking. Keep an eye out for unusual data volumes or questionable data transfers in outbound network traffic.

- Secure endpoints (such as laptops and mobile devices) to stop data leaks from equipment that leaves the organization's grounds. Implement safety measures including endpoint security programmers, remote data wiping, and full disc encryption.

- Employee education and information, employees should receive regular training on the dangers of data leaking, the value of data protection, and the best ways to handle sensitive information. the potential repercussions of data leaks and social engineering techniques.

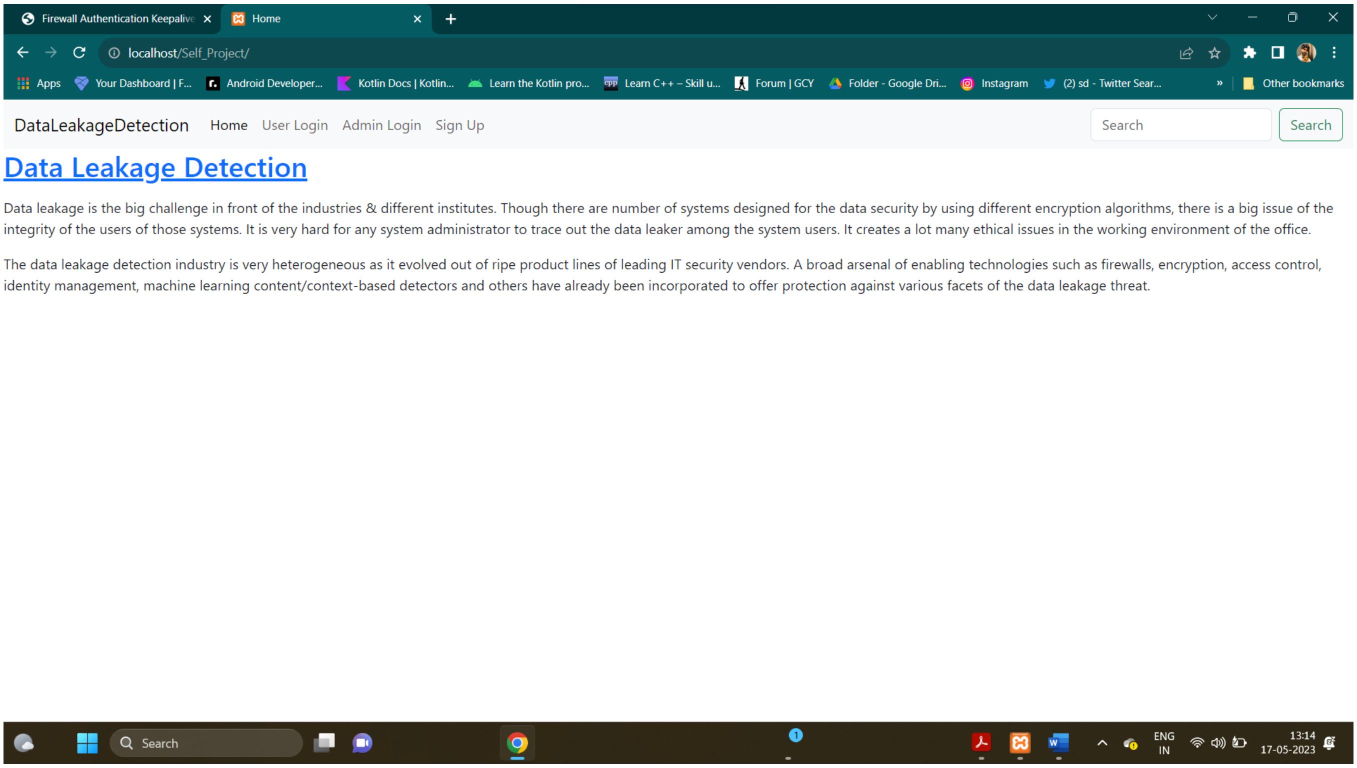
- Create an incident response strategy that explains what should be done in the event of a data leakage incident. Include protocols for inquiry, containment, notification, and recovery.

- Conduct regular audits and inspections of your data leakage detection system to make sure it is working properly. based on the changing threat landscape, identify vulnerabilities, review security measures, and update policies and procedures.

- Continuous development, keep abreast of new developments in the field of data leak detection and prevention. Review and improve your system frequently to address new threats and weaknesses.

To apply solutions in a data leakage detection system, keep in mind that a multi-layered strategy including technical controls, policies, and staff awareness is necessary. To effectively address changing risks and safeguard sensitive data, it is imperative to analyses and modify your system on a frequent basis.

4.2 System Screenshots

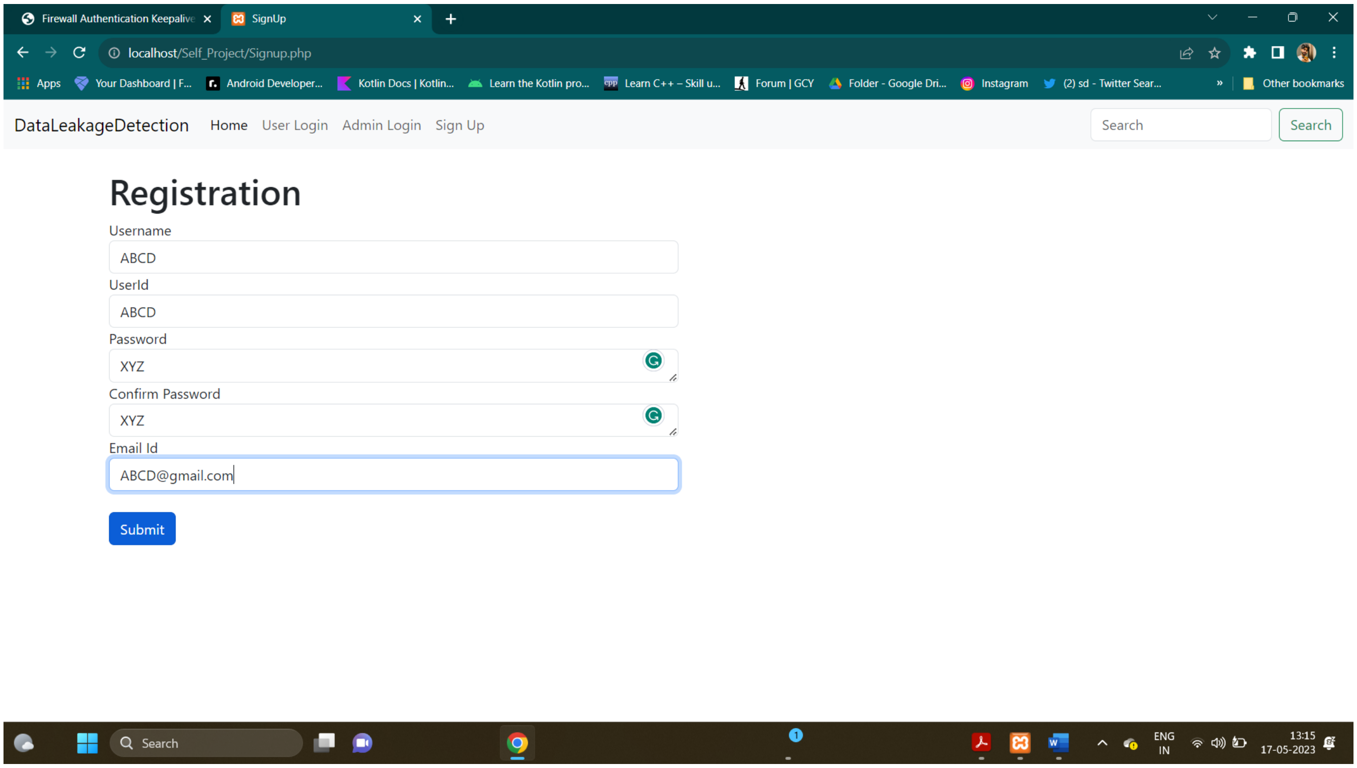


Data Leakage Detection

Data leakage is the big challenge in front of the industries & different institutes. Though there are number of systems designed for the data security by using different encryption algorithms, there is a big issue of the integrity of the users of those systems. It is very hard for any system administrator to trace out the data leaker among the system users. It creates a lot many ethical issues in the working environment of the office.

The data leakage detection industry is very heterogeneous as it evolved out of ripe product lines of leading IT security vendors. A broad arsenal of enabling technologies such as firewalls, encryption, access control, identity management, machine learning content/context-based detectors and others have already been incorporated to offer protection against various facets of the data leakage threat.

Fig 4.1- Index Page



Registration

Username
ABCD

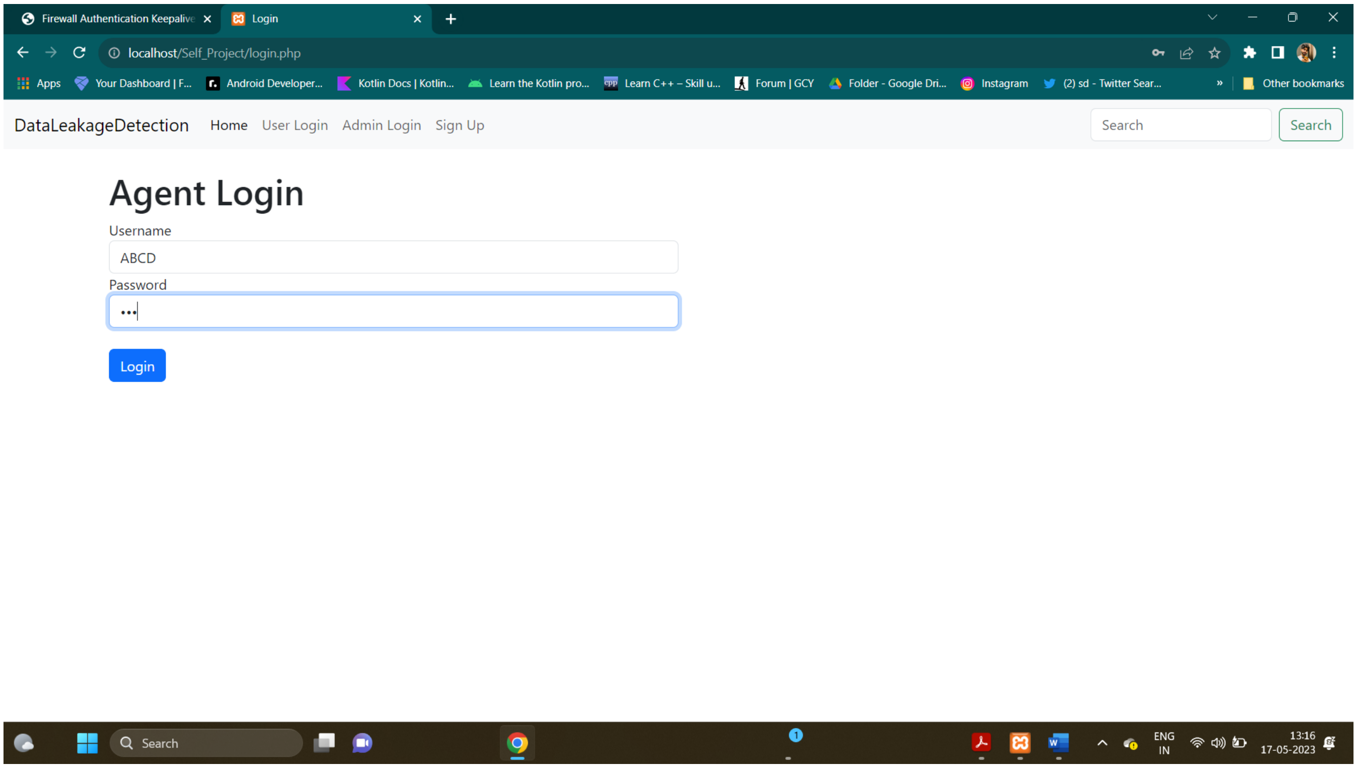
UserId
ABCD

Password
XYZ

Confirm Password
XYZ

Email Id
ABCD@gmail.com

Fig 4.2- Registration Page

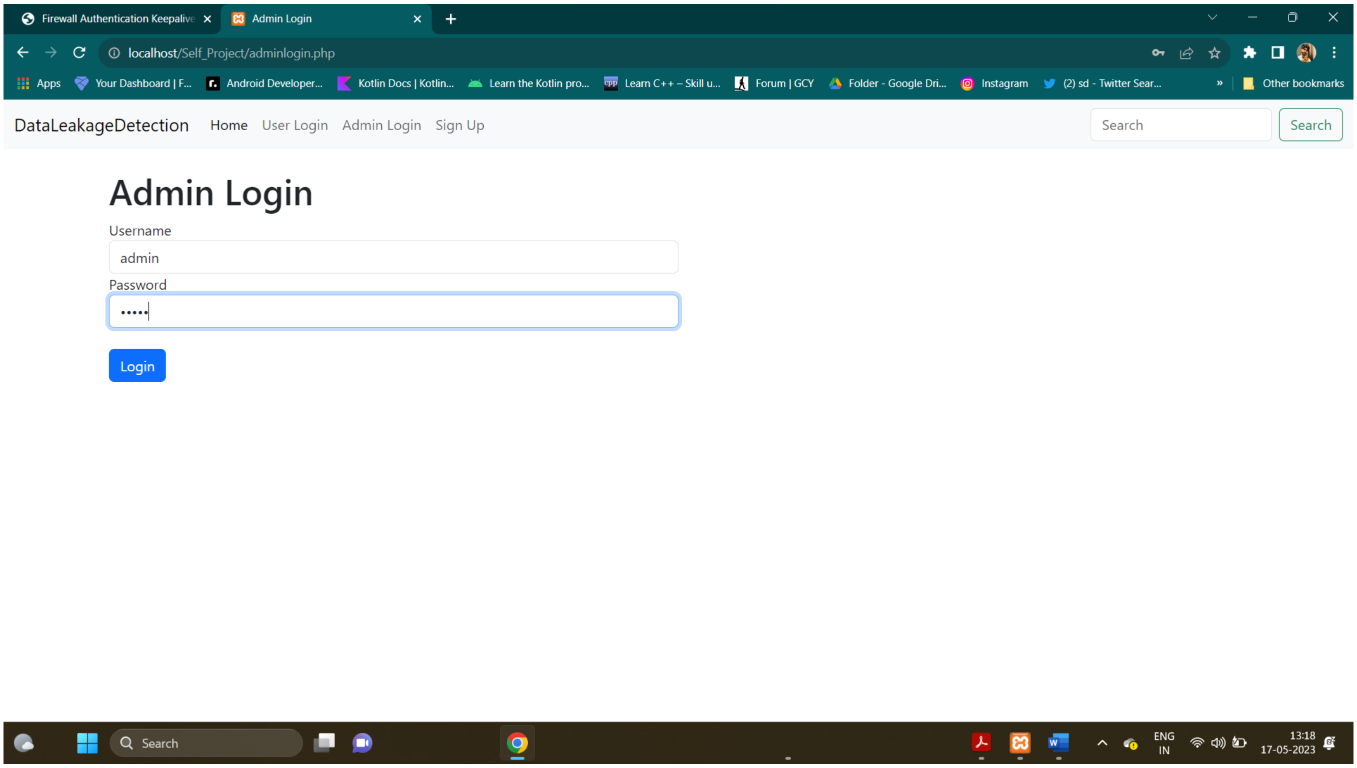


Agent Login

Username
ABCD

Password
•••

Fig 4.3- Agent Login Page



Admin Login

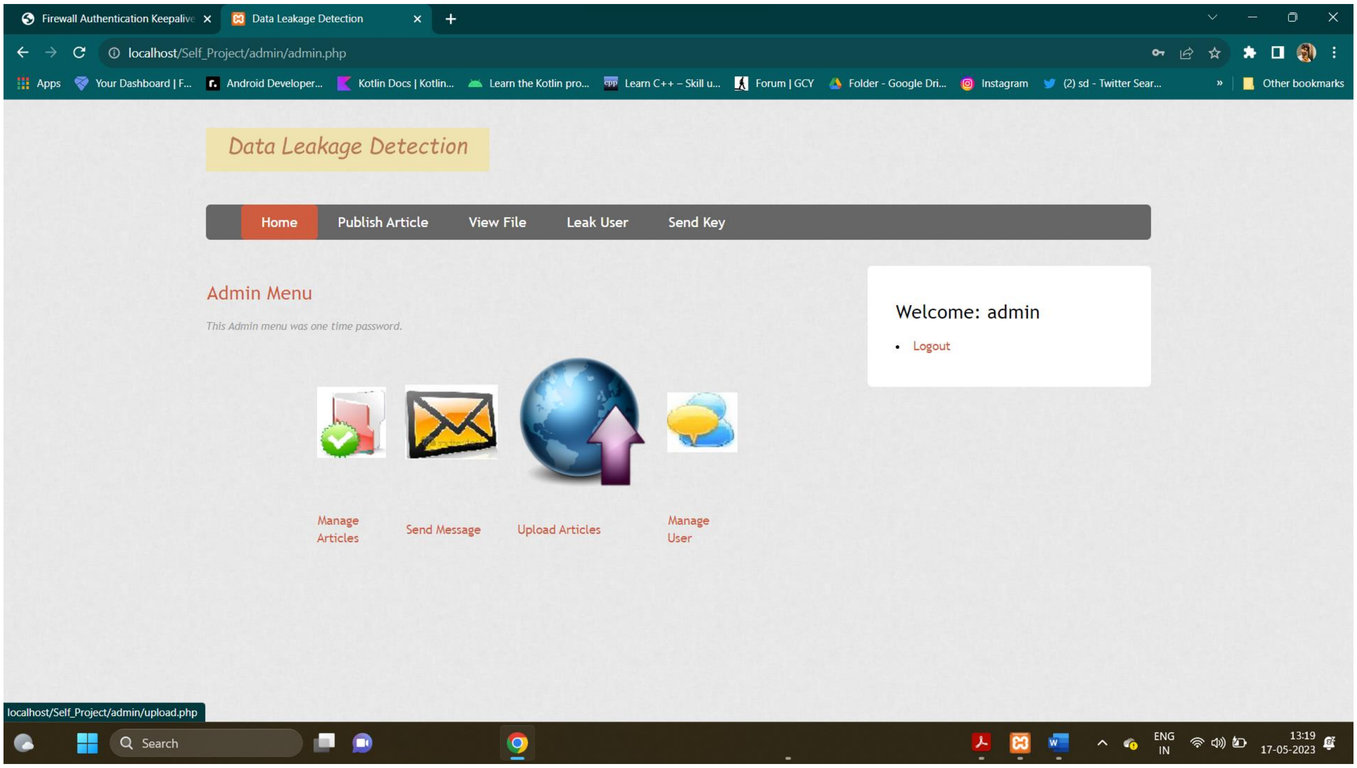
Username

admin

Password

•••••

Fig 4.4- Admin Login Page

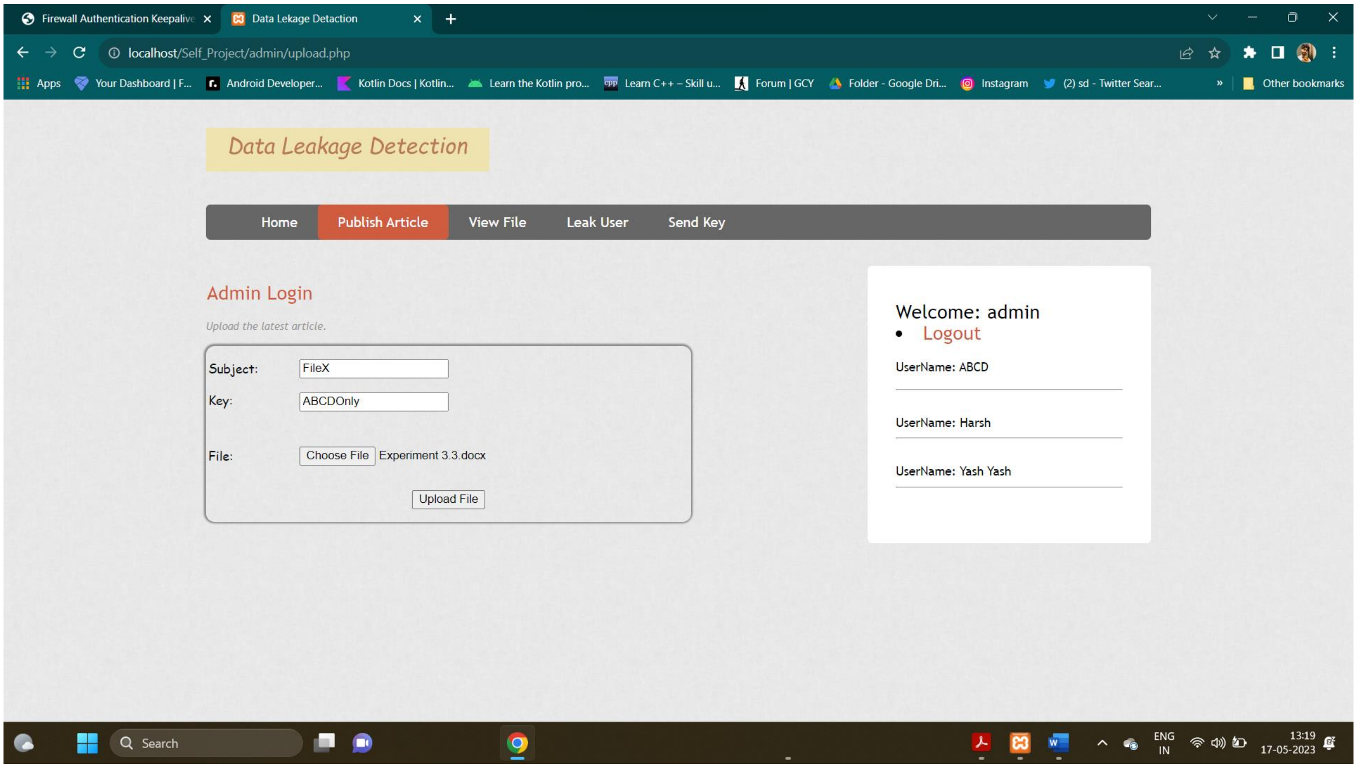


Admin Menu

This Admin menu was one time password.

Welcome: admin

Fig 4.5- Admin Home Page



Data Leakage Detection

Admin Login

Upload the latest article.

Subject:	FileX
Key:	ABCDOnly
File:	Choose File Experiment 3.3.docx

Upload File

Welcome: admin

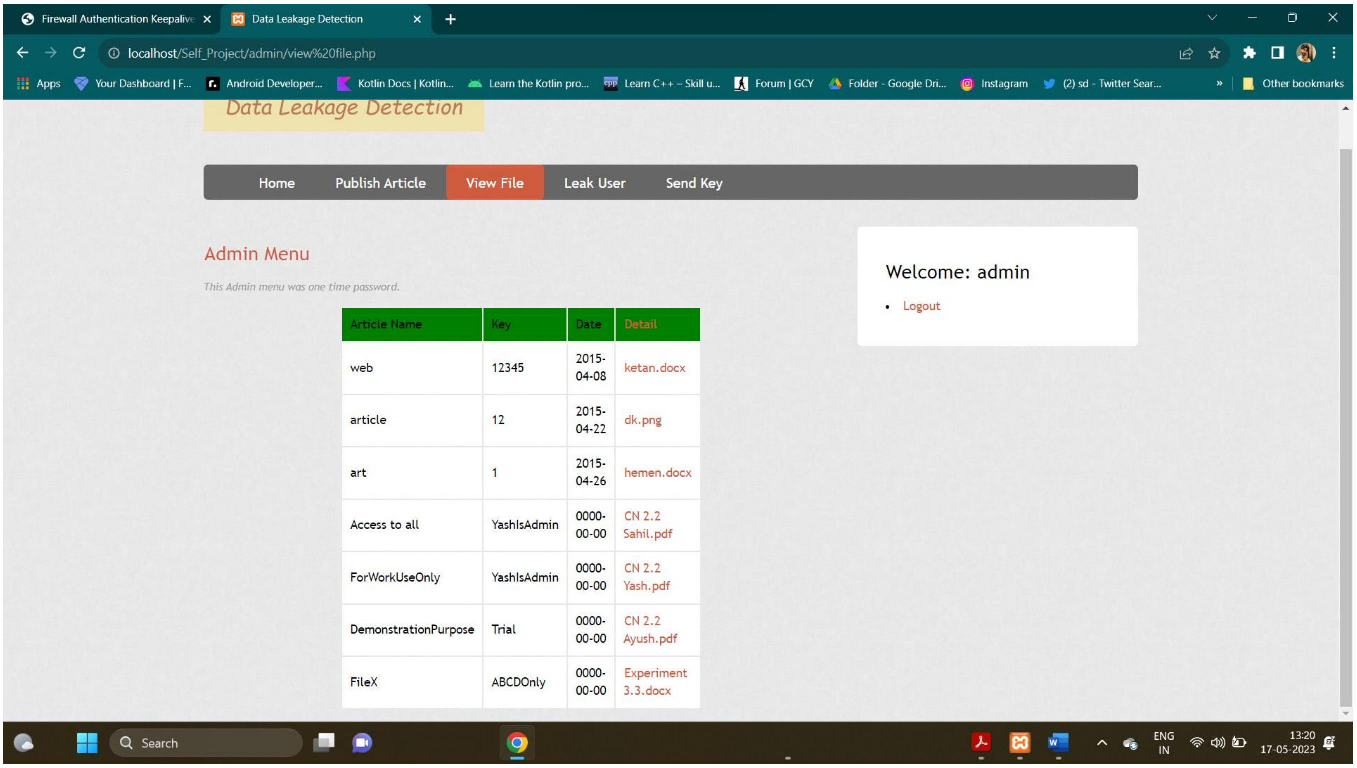
- Logout

UserName: ABCD

UserName: Harsh

UserName: Yash Yash

Fig 4.6- Admin Page to Upload Files on Server



Data Leakage Detection

Admin Menu

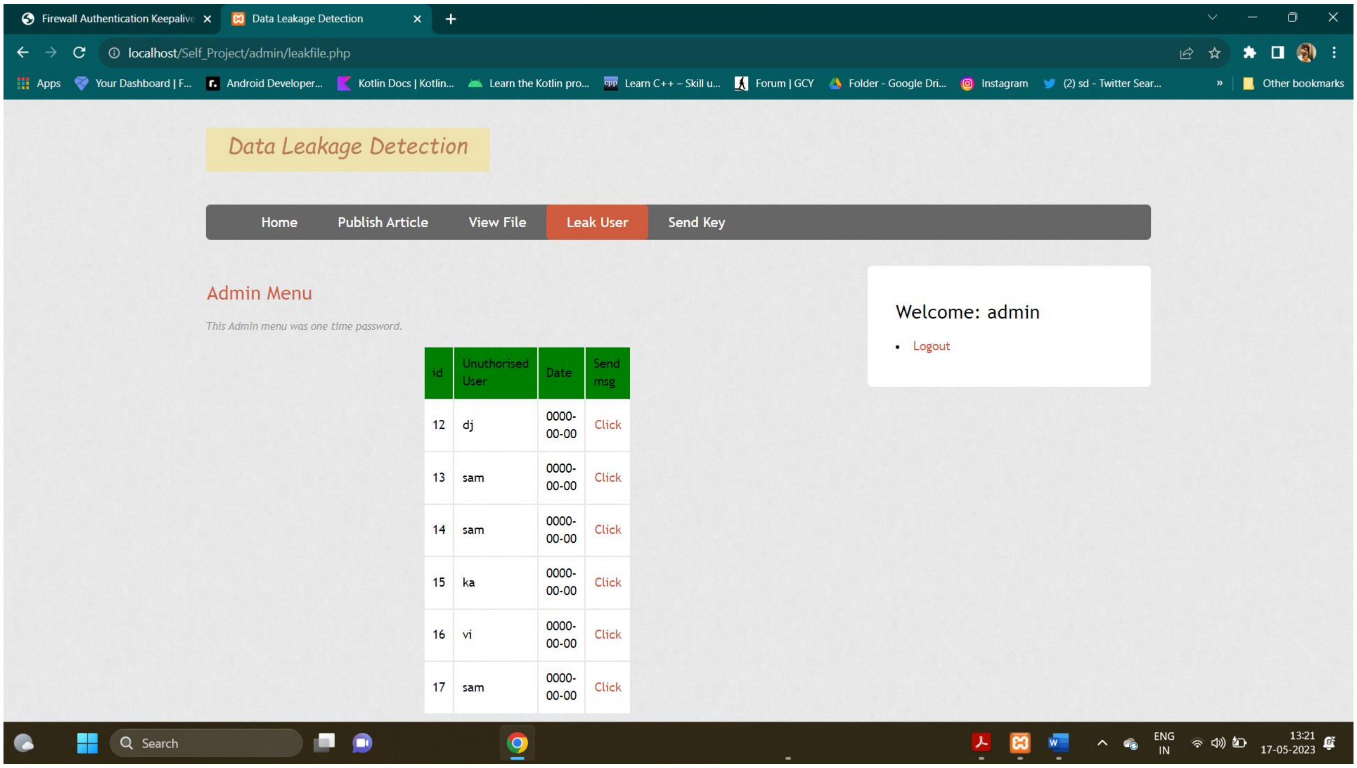
This Admin menu was one time password.

Article Name	Key	Date	Detail
web	12345	2015-04-08	ketan.docx
article	12	2015-04-22	dk.png
art	1	2015-04-26	hemen.docx
Access to all	YashIsAdmin	0000-00-00	CN 2.2 Sahil.pdf
ForWorkUseOnly	YashIsAdmin	0000-00-00	CN 2.2 Yash.pdf
DemonstrationPurpose	Trial	0000-00-00	CN 2.2 Ayush.pdf
FileX	ABCDOnly	0000-00-00	Experiment 3.3.docx

Welcome: admin

- Logout

Fig 4.7- Admin Page to View Files



Data Leakage Detection

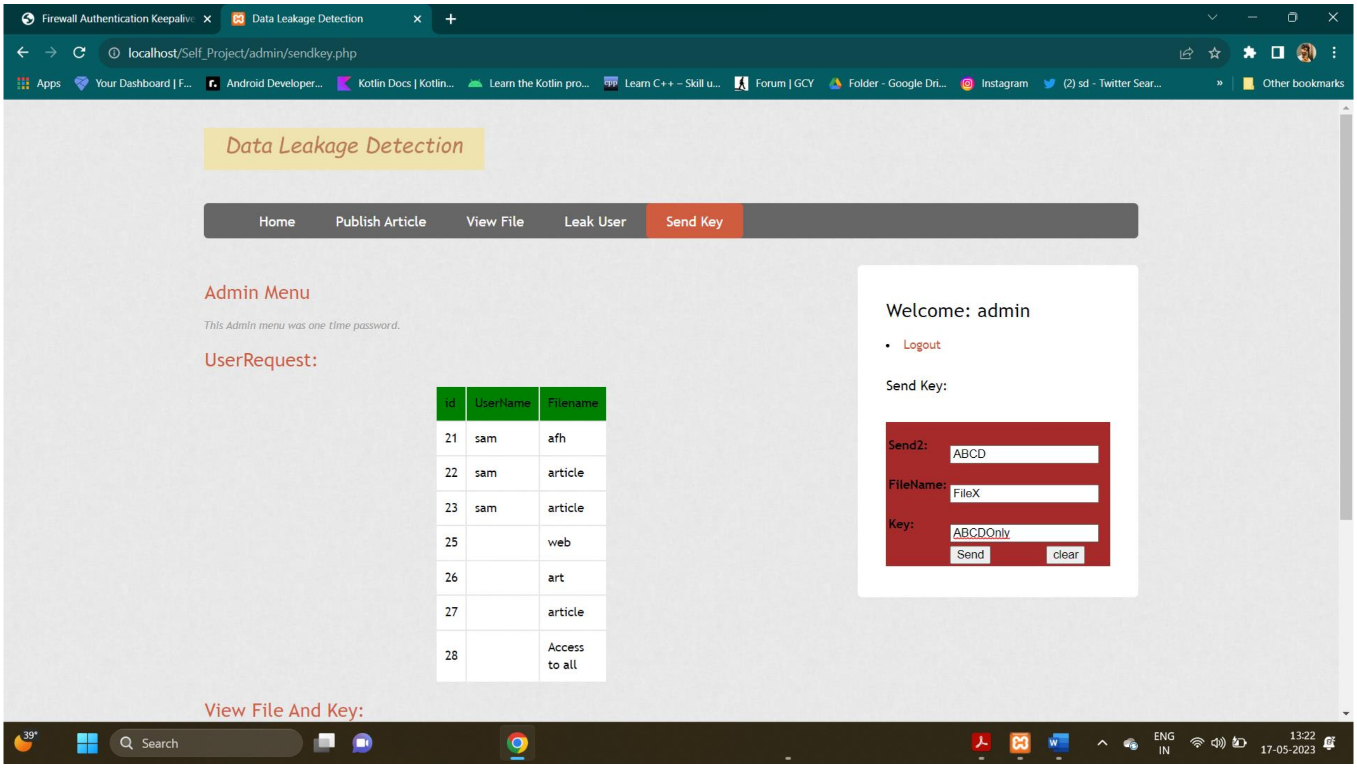
Admin Menu

This Admin menu was one time password.

Welcome: admin

id	Unuthorised User	Date	Send msg
12	dj	0000-00-00	Click
13	sam	0000-00-00	Click
14	sam	0000-00-00	Click
15	ka	0000-00-00	Click
16	vi	0000-00-00	Click
17	sam	0000-00-00	Click

Fig 4.8- Admin Page to Send Unauthorized Users warning message



Data Leakage Detection

Admin Menu

This Admin menu was one time password.

UserRequest:

id	UserName	Filename
21	sam	afh
22	sam	article
23	sam	article
25		web
26		art
27		article
28		Access to all

View File And Key:

Welcome: admin

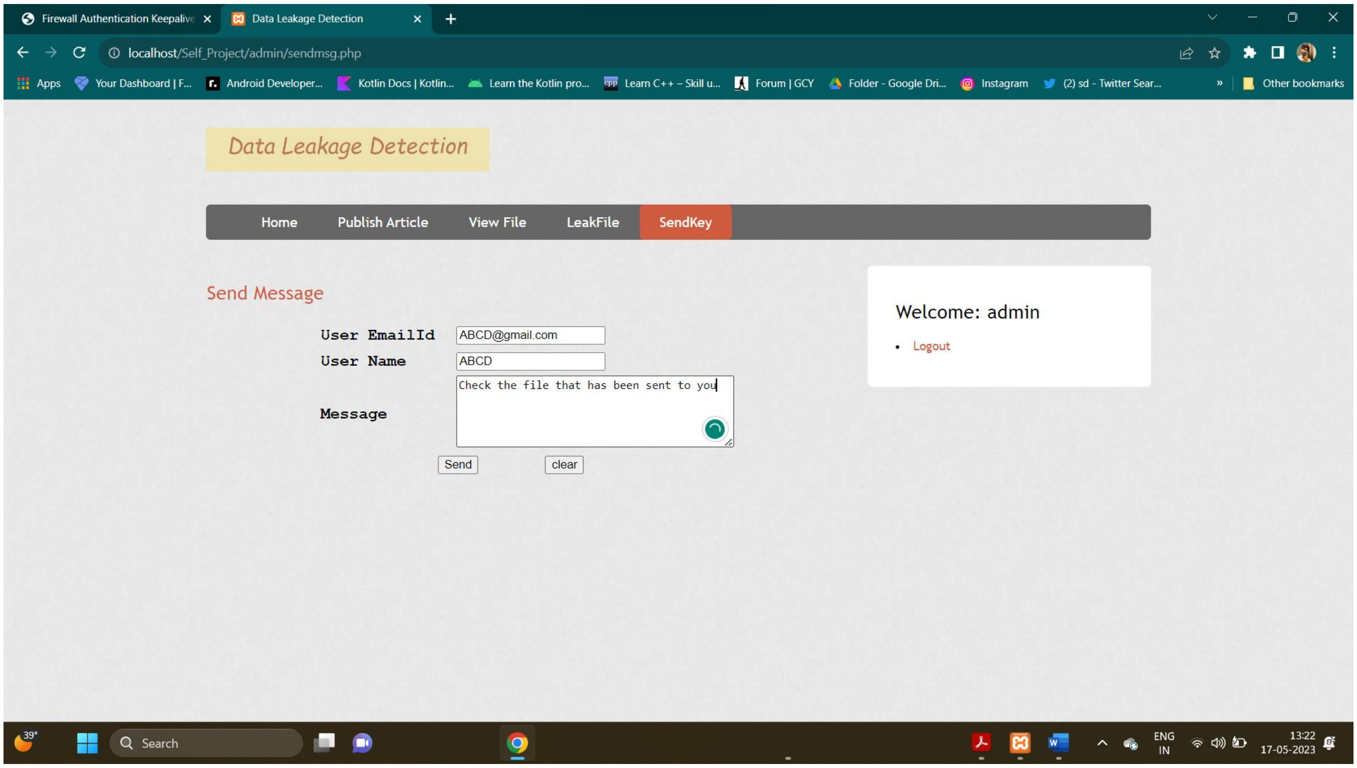
Send Key:

Send2: ABCD

FileName: FileX

Key: ABCDOnly

Fig 4.9- Admin Page To Send Keys to Requested Users



Data Leakage Detection

Send Message

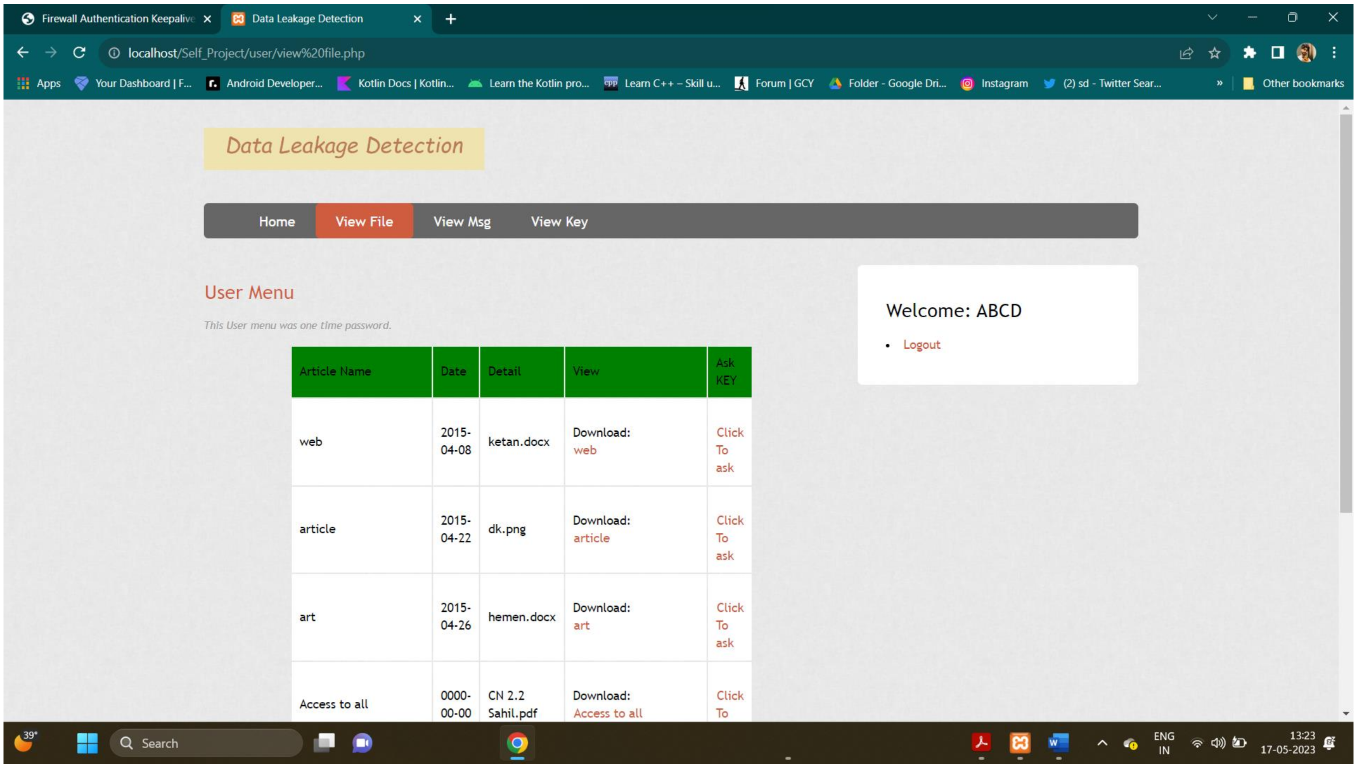
User EmailId: ABCD@gmail.com

User Name: ABCD

Message: Check the file that has been sent to you

Welcome: admin

Fig 4.10- Admin Page to Send Messages to the Users



Data Leakage Detection

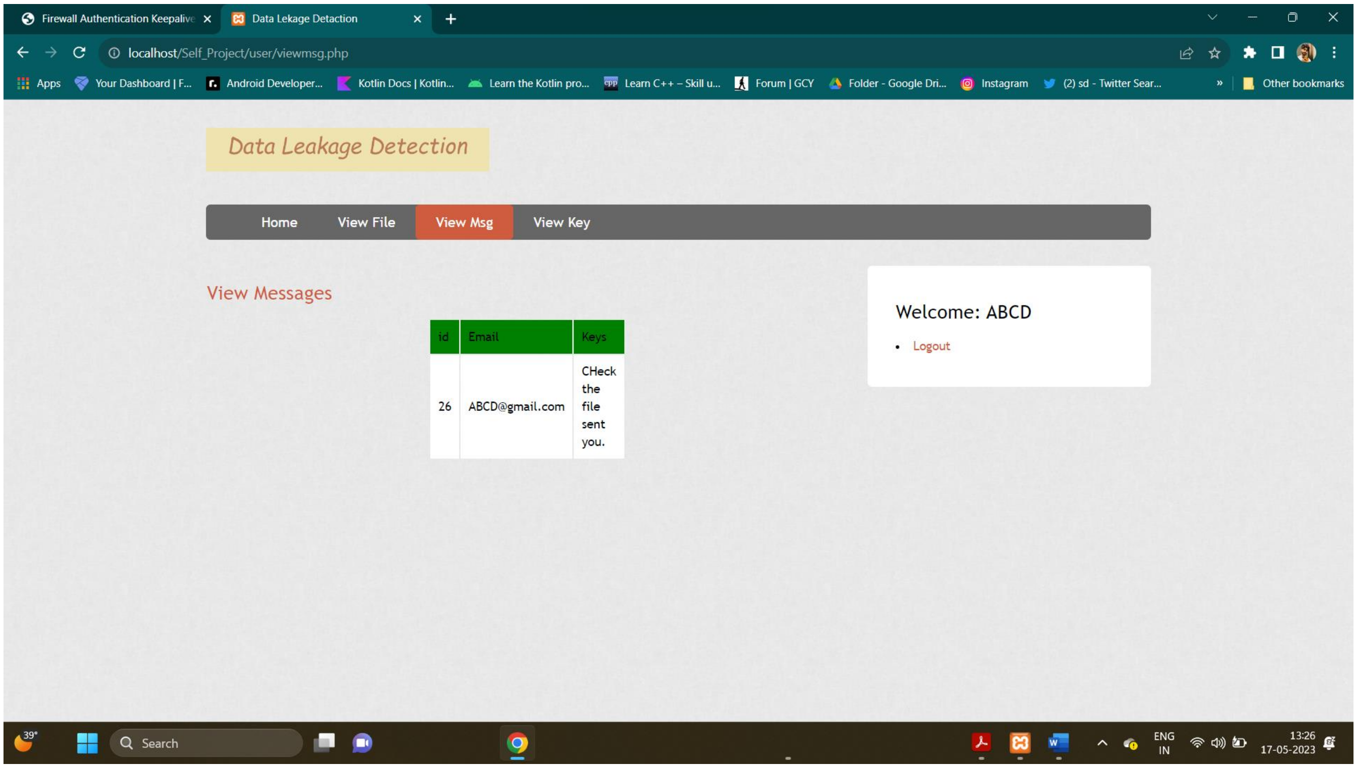
User Menu

This User menu was one time password.

Welcome: ABCD

Article Name	Date	Detail	View	Ask KEY
web	2015-04-08	ketan.docx	Download: web	Click To ask
article	2015-04-22	dk.png	Download: article	Click To ask
art	2015-04-26	hemen.docx	Download: art	Click To ask
Access to all	0000-00-00	CN 2.2 Sahil.pdf	Download: Access to all	Click To

Fig 4.11- User Page to View Files Available on the System



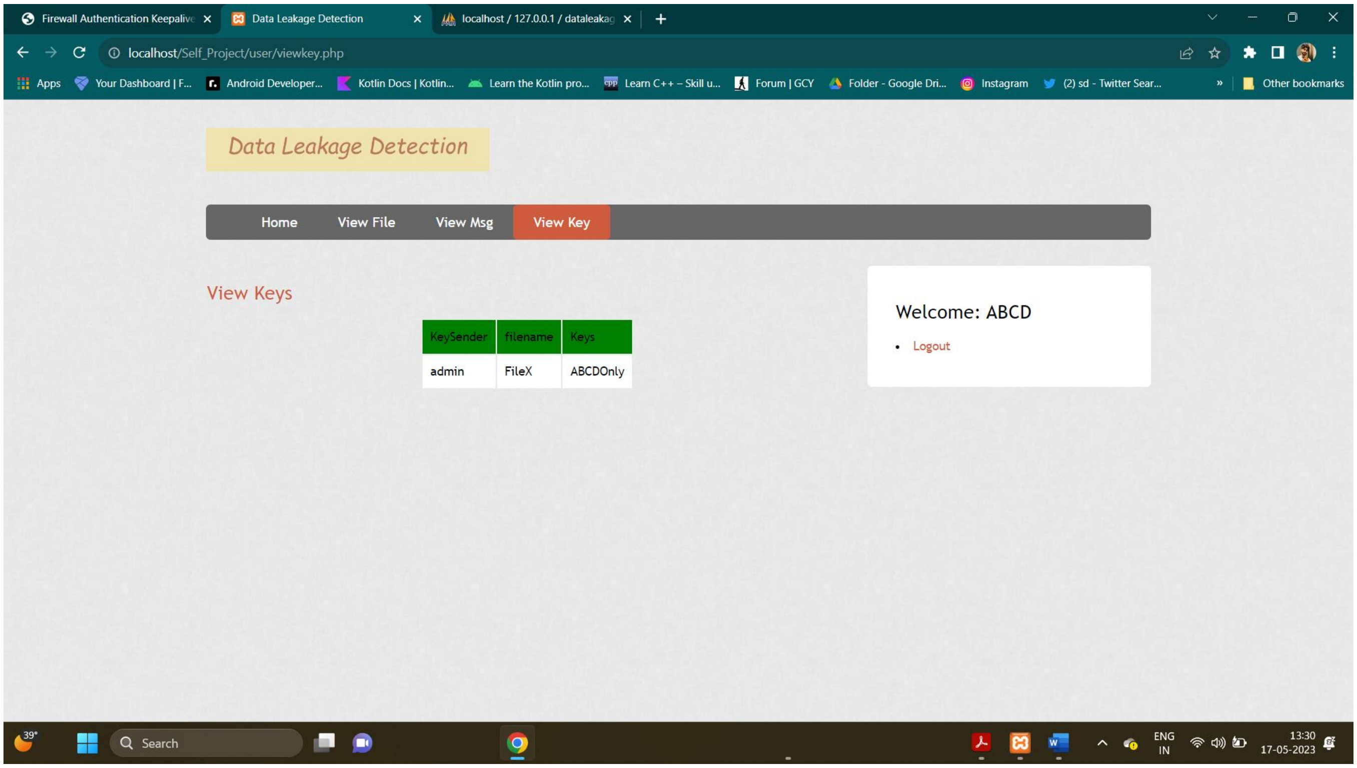
View Messages

id	Email	Keys
26	ABCD@gmail.com	CHeck the file sent you.

Welcome: ABCD

- Logout

Fig 4.12- User Page to View Messages Sent by the Admin



View Keys

KeySender	filename	Keys
admin	FileX	ABCDOnly

Welcome: ABCD

- Logout

Fig 4.13- User Page to View Keys Sent by the Admin

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In order to safeguard sensitive data from unauthorized exposure or leaking, organizations must implement a data leakage detection system. Organizations may greatly lower the risk of data leaks and prevent the potential damage brought on by such accidents by putting in place a comprehensive set of remedies.

Establishing a data leakage policy, categorizing data according to its sensitivity, implementing access controls and encryption, deploying Data Loss Prevention (DLP) tools, monitoring user activity and network traffic, securing endpoints, providing employee training and awareness, creating an incident response plan, carrying out routine audits, and continuously improving the system are all essential parts of a data leakage detection system.

Together, these steps build a multi-layered barrier against data leakage occurrences. Organizations can prevent unauthorized access to sensitive information by setting explicit policies, classifying data appropriately, and putting in place technical controls. DLP solutions, network traffic analysis, and user activity monitoring all work together to quickly identify and stop data leaks. Data held on devices that are beyond the walls of the organization are protected by endpoint security measures.

In order to ensure that staff understand their roles and adhere to recommended practices to prevent data leaks, employee training and awareness programmers are essential. Continuous improvement guarantees that the system adapts to address new threats and vulnerabilities while regular audits and assessments help uncover weaknesses and guarantee the effectiveness of the system.

Sensitive information is protected by a well-implemented data leakage detection system, which also helps an organization stay in compliance with legal obligations and builds stakeholders' trust.

A data leakage detection system, in general, is a proactive strategy to safeguard confidential information, reduce the likelihood of data leaks, and lessen the potential financial, legal, and reputational repercussions connected with data breaches.

5.2 Future Work

Future technological developments and shifting threat environments will continue to influence how data leakage detection systems are created and enhanced. Here are some considerable research areas:

- Advanced machine learning and AI methods: Machine learning and AI advancements can be used to increase the precision and potency of data leakage detection systems. Deep learning, natural language processing, and anomaly detection techniques can be used to more correctly analyze data patterns, spot anomalous activities, and find probable data leaks. These methods can lower false positives and raise the detection of data leakage's overall effectiveness.

- Behavior-based anomaly detection: Behavior-based anomaly detection can be used to complement conventional rule-based methods for detecting data leakage. Organizations can find suspicious behaviors that might point to data leakage by creating baseline user behavior profiles and regularly looking for departures from these patterns. Subtle deviations or abnormal user behavior that rule-based systems can miss can be found with the use of advanced analytics and machine learning techniques.

- Integration of threat intelligence: Real-time threat intelligence feeds can offer insightful details on new threats, attack strategies, and indicators of compromise. The ability of data leakage detection systems to recognize and stop cutting-edge data leakage techniques can be improved by integrating such threat intelligence. Organizations can proactively modify their detection techniques and policies to address new risks by keeping up with the most recent threat landscape.

- Support for cloud and hybrid settings: As businesses increasingly utilize cloud and hybrid environments, data leakage detection systems in the future should be built to work with these intricate infrastructures. Data transfers and storage between various cloud providers and on-premises systems should be thoroughly monitored and protected by these systems. Future data leakage detection systems must take into account factors including ensuring data security in multi-cloud environments, integrating with cloud security services, and putting in place reliable encryption and access controls.

- User-centric strategies: Recognizing user behavior and intents can help data leakage detection systems work more efficiently. User behaviors, access patterns, and contextual data are analyzed as part of user-centric approaches to spot potential insider threats or malicious behavior. Risk scoring, user profile, and behavior analysis can all be used to find high-risk users or accounts that may be more likely to leak data. Organizations can more effectively identify and counter insider threats by taking the human dimension into account while preventing data leakage.

- Integration with threat hunting and incident response: Data leakage detection systems can be strengthened by integrating with threat hunting tools and incident response procedures. Organizations can actively look for potential data leaks and signs of compromise by fusing data leakage detection with proactive threat hunting efforts. Organizations can quickly recognize, contain, and lessen the effects of data leakage incidents thanks to this proactive approach and effective incident response systems.

- Technologies that protect personal information: With the emergence of privacy laws like the General Data Protection Regulation (GDPR), privacy is a crucial factor in data leakage detection systems. Systems in the future should incorporate privacy-enhancing technology to guarantee compliance and safeguard sensitive data. Differential privacy, secure multi-party computation, and homomorphic encryption are some of the techniques that can be used to protect data privacy while still providing efficient data leak detection and prevention.

- Monitoring across platforms and communication channels: Systems for detecting data leaks should be able to monitor across a variety of operating systems, software programs, and communication routes. This includes keeping track of data transfers through social media, instant messaging apps, collaboration software, and other newer methods of communication. Organizations can provide thorough coverage and guard against data leaks that take place over a number of channels by monitoring data flows across numerous platforms and channels.

- Future data leakage detection systems should put a strong emphasis on continuous monitoring and real-time response capabilities. To promptly identify and react to data leakage problems, continuous monitoring entails tracking data flows, system activity, and user behaviors in real-time. Automated warnings and other real-time reaction mechanisms can help businesses respond quickly to data leaks and successfully lessen their effects.

- Prioritization based on risk: Using a risk-based strategy can assist focus resources and work toward finding data leaks. Organizations can better spend their monitoring and preventive efforts by determining the risk connected to various data assets and users. In contrast to lower-risk areas, higher-risk data assets or users may be subject to more intense surveillance and stringent regulations. This strategy maximizes the effectiveness of the detection system by allocating resources to regions where data leakage may have a more negative impact.

- Data leakage detection systems can profit from interaction with more comprehensive security analytics platforms. Organizations can get a complete picture of their security environment by integrating data leakage detection with other security tools and data sources, such as intrusion detection systems, log management systems, and threat intelligence platforms. This connection enables a more comprehensive approach to the detection and prevention of data leakage and allows for improved correlation and analysis of security incidents.

- Automation and orchestration: These tools can speed up the processes for identifying and responding to data leaks. To handle rote duties like incident triaging, data classification, and response coordination, automated workflows can be set up. Through the automated sharing of information and response actions made possible by orchestration tools, incidents can be resolved more quickly and effectively.

- Contextual analysis: Adding contextual analysis to data leakage detection systems can give users broader understanding of the type and scope of potential data leaks. Organizations can gain a better understanding of the context in which data leakage incidents take place by evaluating contextual data such as user roles, data access patterns, geographic locations, and timeframes. This data can support incident investigations, aid in incident prioritization, and increase the overall accuracy of data leakage detection.

- Continuous evaluation of encryption and data protection technologies: As encryption and data protection technologies advance, it is essential to continuously evaluate their efficacy. Future data leakage detection systems should have controls to check that encryption and data security techniques are being used correctly and with integrity. This guarantees that sensitive data is kept appropriately safeguarded and that any potential flaws or incorrect setups are found and fixed right away.

- Integrating data leakage detection systems with identity and access management (IAM) programs can improve security measures and stop unwanted access to confidential information. Organizations may enforce granular permissions, manage user lifecycles, and stop unwanted data access by utilizing IAM capabilities including user provisioning, authentication, and access restrictions. This integration improves the overall approach for preventing data leaks and reduces the danger of data breaches brought on by hacked user credentials or access permissions.

In conclusion, utilizing cutting-edge technology, including user-centric strategies, integrating with other security tools and platforms, and adopting automation and contextual analysis are key to the future of data leakage detection systems. Organizations can improve the efficacy, accuracy, and efficiency of their data leakage detection efforts and more effectively safeguard their sensitive data from unauthorized disclosure or leaking by concentrating on these areas.

REFERENCES

1. Kim, D. H., Lee, H. J., & Park, J. H. (2019). Anomaly-based intrusion detection in data leakage prevention. Journal of Supercomputing, 75(9), 5441-5455.

2. Li, S., Wen, Q., Wang, L., & Li, Q. (2018). Machine learning-based data leakage detection for cloud computing. IEEE Transactions on Cloud Computing, 6(2), 358-370.

3. Zhang, L., Chen, Y., Li, P., Xie, X., & Wang, X. (2020). Deep learning-based data leakage detection in mobile environments. Future Generation Computer Systems, 102, 49-57.

4. Alshammari, R., & Elleithy, K. (2019). A review of data leakage detection systems. Journal of King Saud University-Computer and Information Sciences, 31(2), 206-221.

5. Amrehn, M., & Müller, G. (2018, July). Towards a systematic review of data leakage detection. In 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC) (pp. 319-328). IEEE.

6. Kumar, A., & Kumar, P. (2019). Review of Data Leakage Detection Techniques. Journal of Information Security, 10(1), 1-12.

7. Srivastava, S., & Gupta, B. (2021). Data leakage detection systems: A systematic literature review. Journal of Information Security and Applications, 60, 102733.

8. Wu, X., Zhang, Y., Xu, X., & Lu, J. (2019). A comprehensive survey on data leakage detection techniques. Future Generation Computer Systems, 92, 107-123.