A Review Paper on ThreatXplain: Explainable Artificial Intelligence For Malicious Threat Detection For Insider Theft And Phishing

Dhanashri Anil Aher[1], Kashish Namdev Badgujar[2], Jay Yogesh Chavan[3], Akash Bhausaheb Aher[4], Prof. Manjusha Gaikwad[5]

[1,2,3,4]UG Scholar
[1,2,3,4,5]Dept of Computer Engineering
MET's, BKC, IOE, Nashik, Maharashtra, India.

Abstract:
In today's digital world, organizations face major cybersecurity threats such as data leakage and phishing attacks, which can result in loss of confidential information, financial damage, and harm to reputation. To address these challenges, this project proposes an intelligent system powered by Machine Learning that can detect and predict possible data leaks and identify phishing websites. The system works by creating and training a dataset using organizational log files. This helps it understand normal and abnormal data usage patterns. Registered users can access a dashboard where they can upload datasets or manually enter data parameters to check for signs of data leakage or authentication issues caused by employees. After analysis, the system generates a detailed report explaining how the leakage was detected and what factors influenced the prediction. Along with data leakage detection, the system also includes a phishing website detection module. In this module, users can enter any URL, and the system analyzes it using Machine Learning algorithms to determine whether the website is legitimate or a phishing attempt. The result includes an explanation that highlights the features and parameters that helped identify the phishing activity. Overall, this project aims to improve organizational cybersecurity by combining data leakage prediction and phishing detection in a single, user-friendly platform that helps organizations stay protected from internal and external digital threats.

Keywords: Explainable AI, SHAP, LIME, data leakage prediction, phishing detection.

1.0 INTRODUCTION

In the modern digital era, organizations depend heavily on online systems, cloud storage, and digital communication to manage their day-to-day operations. While this has made work faster and more efficient, it has also introduced serious security risks such as data leakage and phishing attacks. Sensitive information like employee details, financial data, and business plans can be exposed either intentionally or accidentally, leading to major losses for the organization. Therefore, ensuring data safety has become a top priority for every company. Data leakage happens when confidential information is shared or accessed by unauthorized people, either due to system flaws or insider activities. In many cases, employees may unintentionally leak data through emails, filesharing platforms, or insecure connections.

Traditional security systems often fail to detect such incidents in time. This is where Machine Learning (ML)comes in — it can analyze large amounts of organizational data, learn from patterns, and identify suspicious activities before they cause harm. This project uses Machine Learning technology to build an intelligent system that can detect and predict data leakage within an organization. The system studies the organization's log files to understand normal data usage behavior and detect any unusual activities that might indicate a leak. It also allows users to upload their datasets or manually enter details to check for any possible data leakage or authentication problems. Once the analysis is done, the system provides a detailed report showing how and why the leakage was predicted.In addition to detecting data leakage,

the project also introduces a phishing website detection module. Phishing is one of the most common online scams where fake websites are created to trick users into sharing their personal or financial information. In this module, users can enter any website URL, and the system uses machine learning algorithms to analyze it and determine whether it is genuine or fake. The system also explains the reasons behind its decision, making it easier for users to understand how phishing activities are identified. By combining both data leakage detection and phishing website identification in one platform, this project offers a complete cybersecurity solution for organizations. It helps in identifying internal threats like data misuse and external threats like phishing attacks in real time. The system's easy-to-use dashboard and intelligent analysis make it suit able for businesses of all sizes.Overall, this project aims to enhance data protection and reduce cybersecurity risks by using advanced machine learning techniques. It provides organizations with a smarter, faster, and more efficient way to safeguard their digital assets and maintain trust among clients, employees, and partners.

2.0 LITERATURE REVIEW

In this chapter we will see the various studies and research con- ducted in order to identify the current scenarios and trends in deep learn- Ing and machine learning for Adaptive E-learning system.

2.1. Bhavya Singh Shishodia, Data Leakage Prevention System for Internal Security, 2022 In this paper, Bhavya Singh Shishodia builds a system to watch over company networks and stop secret info from slipping out by mistake or on purpose. It keeps an eye on all the data moving around, spots if someone's trying to grab or send out stuff they shouldn't, and right away warns the IT bosses. Plus, it pulls together reports so teams can see what happened and fix weak spots fast. The upsides are huge for keeping private data safe, especially from sneaky insiders who know the system too well. It helps follow rules and laws about data protection, and those handy reports make audits a breeze. But it can get pricey to set up and run, sometimes yells "danger" when there's nothing there, and needs regular tweaks plus training so everyone uses it right without hassle.

2.2. Suvendu Kumar Nayak, Secure Framework for Data Leakage Detection and Pre vention in IoT Application, 2023 In this paper, Suvendu Kumar Nayak designs a tough shield for smart homes packed with IoT gadgets like cameras and thermostats that chat online. The setup grabs data from these devices, checks it for anything f ishy, locks it down tight before storing or sending, and blocks leaks before they start—think hackers snooping on your live feed. It's a win because it guards per sonal stuff in everyday connected homes, cuts down on leak risks way better than older setups, and keeps talks between devices super secure and reliable.

2.3. Chanchal Patra, A Comparative Study on Detecting Phishing URLs Leveraging Pre-trained BERT Variants, 2024 In this paper, Chanchal Patra rounds up different versions of BERT—an AI whiz at reading text—and pits them against sneaky phishing links that trick you into fake sites to steal info. She breaks down URLs, feeds them into these models, and sees which one nails spotting the bad guys most accurately, like a digital lie detector for web addresses. The cool part is it catches more phonies than regular tools, keeping your logins and data out of crooks' hands with top-notch precision. It even beats out some tried-and-true methods in speed and smarts. That said, it guzzles serious computer muscle to run, the install is a puzzle for non-techies, models need fresh training as tricks evolve, and that whole process eats up hours.

3.0 METHODOLOGY

3.1 Block Diagram

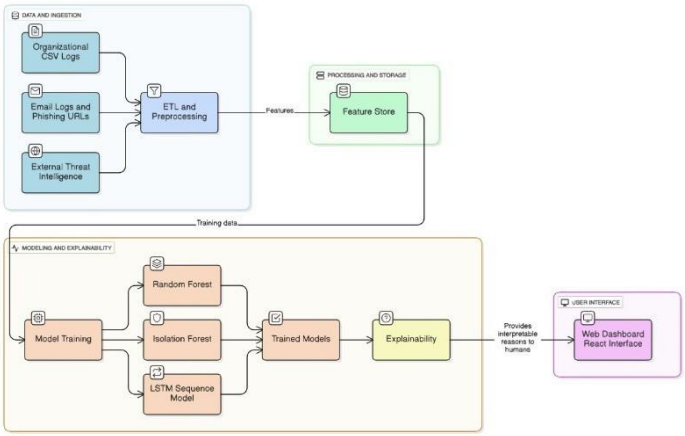


Figure 3.1: Block Diagram

1. **User Login/Registration:** Authorized users log in or register to access the system securely.
2. **Dataset Upload/Input:** User uploads a dataset or provides a URL containing activity logs.
3. **Processing Module:** Handles data cleaning, normalization, feature extraction, and model training.
4. **Dataset Creation & Training:** Data is divided into training and testing sets for accurate model learning.
5. **Model Saving:** The trained model is saved for future threat detection.
6. **Data Leak Detection:** The system detects suspicious activities and possible insider threats.

3.2 Working of the ThreatXplain:

The proposed system ThreatXplain is designed to detect and explain malicious insider threats and phishing activities using machine learning and explainable AI (XAI) techniques.

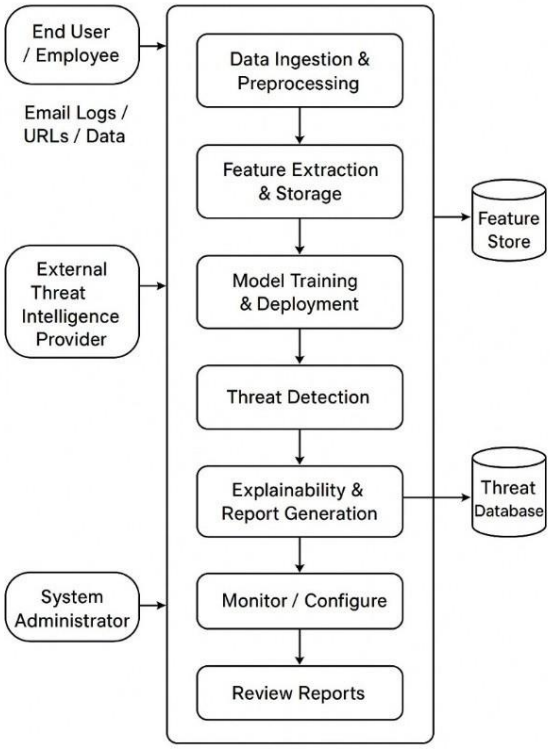
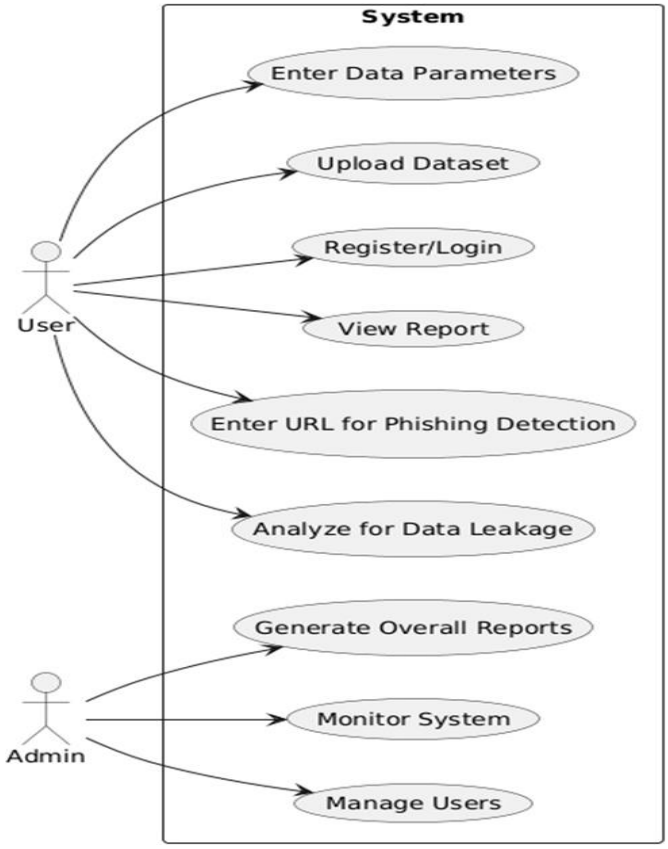


Figure 3.2: Working of the ThreatXplain

1. **Data Collection:** Data is taken from the CERT Insider Threat Dataset, including logs of user logins, file access, emails, web activity, and USB usage for behavior modeling.
2. **Data Preprocessing:** Data is cleaned, normalized, and time-based patterns are extracted using Pandas and NumPy to ensure consistency and structure.
3. **Feature Engineering:** Key features like login time, file usage, and device access are derived to capture behavioral patterns using Scikit-learn.
4. **Detection Engine:** Machine learning models such as Isolation Forest, Autoencoder, and LSTM identify anomalies and potential threats using TensorFlow, PyTorch, and Scikit-learn.
5. **Explainability (XAI):** Techniques like SHAP and LIME explain why a threat was flagged, improving transparency and analyst trust.
6. **Dashboard:** Threat results and risk scores are displayed on a web dashboard built with Flask/Django, React, and Matplotlib.

3.3 UML Diagram:



3.4 PROCESS FLOW

It allows you to specify how your system will accomplish its goals. Activity diagrams show high-level actions chained together to represent a process occurring in your system. An activity diagram is essentially a flowchart, showing flow of control from activity to activity. Unlike a traditional flowchart, an activity diagram shows concurrency as well as branches of control. It focuses on the dynamic flow of a system.

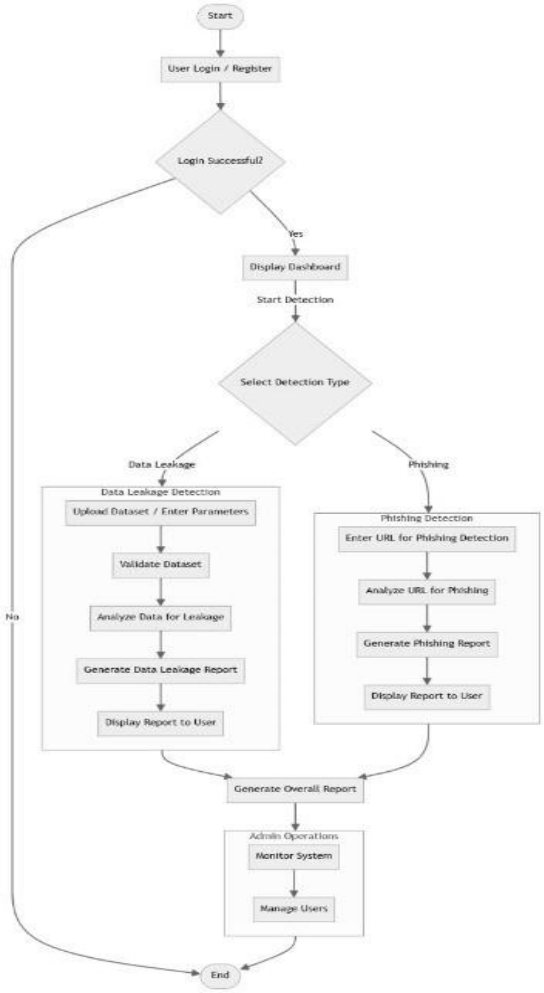


Figure 3.3: Process Flow

4.0 CONCLUSIONS

This project provides an effective and intelligent solution for improving organizational cybersecurity by using Machine Learning to detect data leakage and phishing attacks. It helps identify suspicious activities within an organization and alerts users about possible threats, allowing them to take quick preventive actions. The system's easy-to-use inter face, detailed reports, and accurate predictions make it suitable for both technical and non-technical users. By combining data leakage detection and phishing website analysis in one platform, the project ensures better protection of sensitive information and builds a safer digital environment for organizations.

ACKNOWLEDGEMENT

Authors are very much thankful to management for providing the facilities to conduct the research work in the institution's laboratory. Also thankful to the Science & Technology department for approving our research proposal and providing us the grants to conduct research activities.

REFERENCES:

1. Singh, Bhavya & Nene, Manisha. (2022). Data Leakage Prevention System for Internal Security. 1-6. 10.1109/INCOFT55651.2022.10094509
2. Nayak, Suvendu & Swain, Sangram & Mohanta, Bhabendu & Paikaray, Bijay.(2022). Secure Framework for Data Leakage Detection and Prevention in IoT Application. 1-6. 10.1109 / iSSSC56467.2022.10051336.
3. C. Patra, D. Giri, T. Maitra and B. Kundu, "A Comparative Study on Detecting Phishing URLs Leveraging Pre-trained BERT Variants," 2024 6th International Conference on Computational Intelligence and Networks (CINE), Bhubaneswar, India, 2024, pp. 1-6, doi:

10.1109/CINE63708.2024.10881521.
4. R. Gupta and P. Sharma, ”A comprehensive data leakage detection system using machine learning,” IEEE Transactions on Network and Service Management, vol. 17, no. 4, pp. 2345–2356, 2020.
5. S. Patel and N. Desai, ”Phishing website detection using deep learning techniques,” IEEE Transactions on Cybernetics, vol. 50, no. 3, pp. 1234–1245, 2020.