SIGNATURES

…… ……

STUDENT **DATE**

…… ……..

SUPERVISOR **DATE**

…… ……..

CO-SUPERVISOR **DATE**

…… ……..

Faculty Postgraduate committee chairperson **DATE**

…….. ……..

Dean **DATE**

Please note that this document should not be longer than 15 pages excluding cover page

and contents pages. Delete this statement from your document.

Development of Explainable Phishing Detection Framework and Authentication Risk Modeling Using Metadata Link Graphs and Threat Intelligence

by

Lerato Emmanuel Mgwangqa

Student number: 202224833

Email address: 202224833@spu.ac.za

Degree: BSc Hons Computer Sciences

Department: CSIT

Faculty of Natural and Applied Sciences

Supervisor(s): *Dr Martins A. Arasomwan*

April, 2025

Table of Contents

1. Introduction

Cybersecurity has become one of the most critical challenges in today's digital and interconnected society. The increasing reliance on online systems for communication, finance, and infrastructure has made both individuals and organizations vulnerable to an escalating array of cyber threats. Among these, phishing remains one of the most widespread and damaging attack vectors. Phishing attacks have escalated dramatically in recent years, both in volume and sophistication. A 2024 industry analysis reported a 202% increase in phishing email volume and a 703% rise in credential-phishing incidents during the latter half of that year(Phishing Intelligence Report, 2024.). These attacks now span multiple channels (email, collaboration tools, SMS, social media) and increasingly employ AI-generated content and real-time evasion tactics that defeat static defencesPhishing Intelligence Report, (2024). According to IBM Corporation, (2024), the average cost of a phishing-related data breach has soared to $4.76 million per incident, underscoring the need for more robust detection and response systems.

Traditional cybersecurity approaches such as signature-based or heuristic detection were once effective in combating relatively simple, static attacks. However, threat actors have evolved. Modern phishing attacks often use AI-generated content, personalized social engineering, and real-time evasion tactics that bypass outdated filters and static blacklists (Gartner, 2024). This evolution in threat sophistication has rendered many older detection methods inadequate.

The growing sophistication of phishing attacks, paired with their devastating financial and reputational impacts, calls for smarter and more adaptive detection techniques. As attackers become more agile, cybersecurity solutions must not only match this pace but also provide clear, interpretable outputs to those responsible for making security decisions. This is especially important for Security Operations Center (SOC) analysts, who are tasked with rapidly verifying and acting upon alerts. Detection systems that lack transparency can introduce delays, reduce trust, and contribute to alert fatigue. Therefore, enhancing both detection accuracy and explainability has become a key priority for modern cybersecurity research and practice.

Efforts to improve phishing detection have produced a range of advancements. Machine learning models including Random Forests, LSTMs, and even GNNs have demonstrated high detection accuracy across multiple datasets (Elkouay et al., 2024; Thakur et al., 2023) . Simultaneously, the field of Cyber Threat Intelligence (CTI) has gained traction, with platforms like VirusTotal providing real-time threat context and indicators of compromise (Bardakis, 2024) . Explainable AI (XAI) tools like SHAP have also emerged to address the transparency issue in ML systems (FRANK XAVIER GEARHART, 2024; Prity et al., 2024)

Despite these advancements, many of these tools and models are deployed in isolation, limiting their effectiveness when used in live operational settings. Moreover, the detection systems often operate as "black boxes" that offer little insight into why certain emails are flagged as malicious, leaving SOC analysts without meaningful decision support.

A major limitation of current phishing detection frameworks is the fragmentation of components: detection algorithms, CTI feeds, and interpretability mechanisms often function independently, leading to missed connections and reduced effectiveness. Another overlooked area is the post-phishing behavior of users, especially how authentication patterns change after phishing emails are received. While studies such as have shown that anomalies in login behavior often follow successful phishing attempts, most detection frameworks fail to incorporate this information into risk assessments. As a result, many systems are reactive rather than proactive, lacking the ability to adapt to threats in real-time or provide actionable context to analysts.

This research aims to build a unified, explainable AI framework that strengthens phishing detection and authentication risk modeling by bringing together several powerful components

into one system. At its core, the framework will use graph-based metadata analysis leveraging tools like NetworkX to reveal hidden connections between senders, URLs, and domains (Elkouay et al., 2024) . It will also incorporate real-time cyber threat intelligence (CTI) through platforms like VirusTotal to assess the risk level of domains and links on the fly. In addition, it will model user behavior after phishing incidents, flagging suspicious login activity such as geo-location mismatches or unusual login times (Wang et al., 2024) . A Random Forest classifier will power the phishing detection, while SHAP will be used to generate clear, human readable explanations that help analysts understand and trust the system's decisions.

The main goals of this project are to firstly design and test a hybrid model that combines graph features, metadata, and authentication logs secondly apply SHAP to improve transparency and build trust among Security Operations Center (SOC) analysts; and thirdly to design and simulate a lightweight, serverless prototype capable of real-time inference and threat intelligence enrichment. The expected impact is wide-reaching: security teams will gain faster, clearer insights during incidents; small and medium sized businesses will benefit from an affordable and scalable phishing detection tool; and researchers will have a working reference for blending explainability and CTI into real-world systems. By addressing the disconnect between current tools and adding behavioral insight to the mix, this project hopes to contribute to more intelligent, transparent, and proactive cybersecurity defences.

2. Literature review

2.1 Current State of Research

Phishing remains one of the most persistent and damaging threats in today's cybersecurity landscape. Despite ongoing efforts to secure digital systems, attackers continue to exploit human vulnerabilities through deceptive emails, websites, and links designed to harvest sensitive information. As organizations become more reliant on digital infrastructure, the stakes of phishing-related breaches, both financially and operationally have grown significantly.

Recent developments in artificial intelligence (AI) and machine learning (ML) have revolutionized threat detection, particularly in phishing detection. These technologies offer the potential to automate the identification of complex attack patterns, uncover hidden signals in metadata, and improve response times in security operations centers (SOCs). However, while accuracy in detection has improved, many AI-based systems struggle to provide transparent explanations for their predictions. This lack of interpretability limits analyst trust slows down decision making and undermines the operational value of such tools in real world environments.

Furthermore, modern phishing detection models often operate in isolation separated from other important sources of contextual information, such as threat intelligence feeds or user authentication behavior. This fragmented approach limits a system's capacity to deliver a thorough risk assessment, particularly when assaults occur in numerous phases. This study is motivated by these challenges namely, the need for phishing detection models that are not only accurate but also explainable, the growing importance of real-time threat intelligence, and the currently underutilized potential of behavioral signals following an initial compromise. The following sections explore how the research community has approached these challenges, the progress made so far, and where current systems still fall short.

2.2 Systematic Survey of the Literature

Recent advancements in graph-based phishing detection highlight the value of both interpretable and high-performance models. (Elkouay et al., 2024) introduced a URL Graph-Based Model (URLGBM) that applies random walks and PageRank over URL tokens, achieving a detection

accuracy of 98.98%. This approach demonstrates the effectiveness of lightweight, explainable graph representations. Complementing this, (Guo et al., 2025) employed graph neural networks (GNNs) to model sender–URL relationships, showcasing the predictive power of deep graph structures. This trade-off between model accuracy and interpretability continues to shape the direction of phishing detection research. To address this challenge, the proposed study adopts a more transparent and scalable approach using NetworkX to extract interpretable graph features such as node centrality and sender-recipient clustering, that can be directly mapped to observable phishing behavior patterns..

While deep learning remains popular, its lack of transparency remains a limitation. (Thakur et al., 2023) conducted a systematic review of CNNs and LSTMs in phishing detection, emphasizing their black-box nature. In contrast, (Ali et al., 2023) achieved scalable results with metadata-only models, reinforcing the practical value of simpler, explainable approaches like Random Forests. Similarly, more complex hybrid models like BGL-PhishNet (Remya et al., 2025), which combine BERT embeddings, GNNs, and LightGBM, tend to improve accuracy but at the cost of transparency and computational efficiency.

To address explainability, several researchers advocate for XAI tools such as SHAP and LIME. (Gearhart, 2024) found that cybersecurity analysts' trust in AI systems is directly linked to the interpretability of model outputs, supporting the integration of SHAP in detection frameworks. (Prity et al., 2024) and (Al & Al Shwali, n.d.) both applied SHAP and LIME in malware detection and firewall systems, respectively, concluding that such tools significantly improve decision confidence among SOC analysts.

Cyber Threat Intelligence (CTI) has also evolved as a critical component of resilient cybersecurity frameworks. (Bardakis, 2024) emphasized the importance of structured CTI methodologies but noted challenges in data overload and integration. Researchers like (Alturkistani & Chuprat, 2024) are also exploring how large language models (LLMs) can help automate CTI processing. These developments reflect a broader shift toward dynamic, real-time threat enrichment, but many current systems fail to fully capitalize on these possibilities.

An often overlooked component in phishing research is the analysis of post-attack behavior, especially in terms of how authentication patterns shift after a phishing compromise (Authentication behavior modeling). Recent findings show that 68% of security breaches involve human factors, including phishing and the use of stolen credentials (Huisman, n.d.). This underscores the value of integrating phishing detection with authentication anomaly modeling to uncover suspicious login behaviors following a compromise. Similarly, recent research emphasizes the importance of integrating phishing alerts with login anomaly detection to prevent account takeovers. (Zhao et al., 2023) introduced the CEAD framework, which detects compromised email accounts using temporal and spatial login behavior patterns, highlighting the potential of behavioral analysis post-phishing. While some systems like those described in (Maureen Oluchukwuamaka Okafor, 2024) incorporate behavioral biometrics for fraud prevention, they often rely on intrusive data collection, raising ethical and usability concerns in real-time environments

Taken together, these advances point to a growing recognition of the need for detection systems that are not only accurate, but also explainable, context-aware, and behaviourally informed. However, many current models remain siloed focusing on detection without considering post-compromise risk or real-time threat enrichment.

2.3 Limitations of Current Systems

Existing approaches suffer several notable drawbacks.

Lack of transparency: most high-performing models are black-boxes. As Lim *et al.* point out, "Even

though they are effective, these ML based solutions suffer from a serious drawback: explainability. Many phishing detection models work as" black boxes", meaning users and security analysts have little or no idea why an email got classified as phishing. Without transparency, users are less likely to trust the results or warnings and carry out actions."(Lim et al., 2025).

Scalability and performance: complex models (deep networks, graph algorithms or ensembles) can be computationally heavy. For example, a recent review notes that "training and deployment of multiple models are computationally intense, especially when it involves real-time applications"(Kavya & Sumathi, 2024) This makes on-the-fly email filtering or live URL checking challenging at large scale.

Integration gaps: many systems treat phishing detection (at the email/URL layer) separately from post-login risk. In practice, however, a layered pipeline is needed. (Divakaran & Oest, 2022) emphasize that an effective solution often requires "a pipeline of multi-layered approaches, including rules and scores based on other sources". In reality, few research frameworks unify phishing detection with authentication analysis.

Data and adaptability: ML models often rely on features that may not generalize across contexts. (Mia et al., 2025) show that features useful in one phishing dataset may not transfer to another. In RBA, most work is server-centric and does not scale; (Fereidouni et al., 2024). note "limiting automated pattern recognition. Several systems fail to support continuous learning from user context and behavior, hindering model adaptation to new patterns. Additionally, these frameworks often face cold start challenges due to insufficient historical data, resulting in suboptimal security decisions for new users or contexts. To the best of our knowledge, there is no contribution in the open literature that reports the combination of a distributed risk engine for model training essential for user data privacy preservation with improved scalability". In short, current systems lack explainability, struggle with compute/latency, and do not seamlessly combine email-layer intelligence with login-layer risk.

2.4 Motivation for the Proposed Research

The survey above highlights three interrelated gaps in the current research landscape. First, graph-based modeling for phishing detection is often decoupled from real-time threat intelligence and lacks integration with explainability tools like SHAP. Second, authentication behavior is rarely modeled alongside phishing indicators, despite strong evidence of its relevance in assessing account compromise risk. Third, most CTI enrichment approaches remain static and do not provide analysts with actionable, explainable insights.

Considering these gaps, this proposed research aims to develop a unified, explainable framework for phishing detection and authentication risk modeling. By leveraging graph-based metadata analysis, real-time IOC enrichment (e.g., VirusTotal), and post-phishing authentication signals, combined with SHAP explanations the framework seeks to support scalable, analyst-trustworthy decision-making in SOC environments.

Identified Gaps

1. Fragmented Integration of Graph Modeling, Real-Time IOCs, and XAI

- Current State: Recent studies ((Guo et al., 2025); (Al-Sabbagh et al., 2024)) demonstrate graph-based or cluster-based phishing detection. However, they lack seamless integration with real-time threat intelligence (e.g., VirusTotal IOCs) and explainability tools like SHAP, limiting their operational use in SOC environments. Analysts cannot trace dynamic threats transparently.
- Our Solution: Unify graph metrics (e.g., sender-URL centrality) with live IOCs and SHAP-driven rationales.

2. Underutilized Post-Phishing Authentication Behavior as a Risk Signal

- Current State: While (Wang et al., 2024) link phishing to authentication anomalies, most frameworks treat detection and post-compromise behavior as silos. Recent research by (Zhao et al., 2023) underscores the significance of analysing login behaviors to detect compromised email accounts. Their CEAD framework effectively identifies anomalies without relying on labelled data, highlighting a critical gap in current phishing detection methodologies that often overlook the correlation between phishing alerts and subsequent login anomalies
- Our Solution: Integrate post-phishing authentication behaviors (e.g., geo-location mismatches, login frequency shifts) into a unified threat scoring model to support more comprehensive and real-time risk modeling..

3. Static CTI Enrichment Without Explainability

- Current State: CTI frameworks (Bardakis, 2024) focus on data aggregation but lack adaptive enrichment (e.g., updating graph nodes with live IOCs) and contextual explanations.
- Our Solution: Dynamically enrich graph nodes with VirusTotal IOCs and map SHAP values to MITRE ATT&CK tactics for analyst-ready insights.

3. Rationale, Problem statement and hypothesis

The rapid evolution of phishing attacks and the limitations of current detection systems underscore the need for a more integrated and interpretable approach to cybersecurity. Existing solutions, while highly accurate, often operate as opaque “black boxes” and lack seamless integration of real-time threat intelligence and post-phishing authentication behavior analysis. This fragmentation hinders effective decision-making in Security Operations Centers (SOCs) and places organizations at elevated risk for data breaches.

3.1 Problem Statement

Current phishing defence frameworks remain fragmented and reactive. They typically combine isolated components e.g. a spam-filtering classifier here, a threat feed lookup there, a separate MFA alert system but lack a unified, context-aware architecture. This siloed design means emerging phishing tactics (especially AI-generated or polymorphic links) can bypass one layer without tripping others. Furthermore, even when a high-performing model flags an email, SOC analysts often receive only a binary judgment without explanation. In effect, many models boast excellent accuracy on benchmarks but provide no actionable rationale for analysts, making it difficult to validate alerts or prioritize responses (Lim et al., 2025) .

Another critical gap is the omission of post-event authentication risk signals. Research has shown that successful phishing can be followed by anomalies in user login patterns (e.g. sudden login failures, impossible travel between locations, or unexpected MFA challenges). Yet few detection systems correlate an email alert with downstream account activity. The result is a missed opportunity: without linking phishing incidents to user behavior, risk assessments remain incomplete, and response tends to be purely reactive. For example, organizations may detect a suspicious link in an email, but not immediately check whether any credentials have since been compromised. Overall, the challenge is to overcome the brittleness and opacity of existing solutions. We must address a multi-dimensional research gap which is designing an integrated phishing detection framework that is adaptive (leveraging real-time CTI and behavior analytics) and interpretable (providing clear, human-readable justifications) so that SOC teams can make faster, more confident decisions.

3.2 Research rationale

The proposed research is motivated by the need to bridge these gaps by integrating:

- **Graph-Based Modeling:** Utilizing metadata from emails to construct simple, interpretable graph representations that reveal relationships among senders, URLs, and domains (Elkouay et al., 2024).
- **Threat Intelligence Integration:** Enriching these models with real-time Indicators of Compromise (IOCs) from sources like VirusTotal to dynamically validate suspicious behavior (Bardakis, 2024).
- **Authentication Risk Modeling:** Incorporating post-phishing login behavior to assess whether anomalous authentication events—such as geo-location mismatches—correlate with phishing attempts ((Wang et al., 2024).
- **Explainable AI:** Employing SHAP to enhance the transparency of a Random Forest classifier, thereby making the decision-making process accessible and actionable for SOC analysts ((FRANK XAVIER GEARHART, 2024); (Prity et al., 2024)).

3.3 Hypothesis

We hypothesize that a unified framework combining **metadata-driven graph analysis**, **real-time threat intelligence enrichment**, **authentication risk modeling**, and **explainable AI (XAI)** will outperform conventional, isolated systems in both accuracy and operational usability. Specifically, integrating these components should yield markedly higher detection rates (we aim to exceed 90% accuracy on realistic workloads) while also reducing the mean time to resolve incidents. For instance, dynamic graph models can uncover subtle connections among malicious senders and URLs, CTI feeds can flag new threats on the fly, and behavioral analytics can confirm whether a phishing attempt led to account anomalies. Layering SHAP-based explanations over the underlying model will translate these signals into clear rationales (e.g. “this email’s URL was recently observed in a phishing campaign” or “post-click login came from a foreign IP”), thereby empowering analysts. Prior studies in related domains support this approach: graph-based explainable models in fraud detection have simultaneously improved prediction performance and provided interpretable insights for analysts(Li et al., 2024). Likewise, SOC surveys find that providing feature-level explanations and confidence scores greatly enhances triage efficiency(Rastogi et al., 2025). Based on these precedents, we expect our proposed XAI-driven framework to demonstrably improve phishing detection metrics and to significantly accelerate SOC response times, validating the value of the integrated, explainable design.

Delimiting Boundaries:This study focuses exclusively on email-based phishing detection and assumes the availability of structured email data, authentication logs, and access to live threat intelligence APIs. It is designed for resource-constrained environments where lightweight, interpretable models are preferred over computationally expensive deep learning architectures. Although the underlying approach may be extended to other cybersecurity domains, its immediate applicability is limited to traditional email ecosystems.

In summary, the research aims to develop and validate a comprehensive, explainable framework that addresses the current fragmentation in phishing detection systems. By synthesizing state-of-the-art graph modeling, CTI integration, and behavioral risk assessment with SHAP-based interpretability, the project seeks to offer a scalable, transparent, and operationally effective solution tailored for modern SOC environments.

4. Aim and objectives

4.1 Aim

To develop explainable phishing detection framework and authentication risk modeling using metadata link graphs and threat intelligence.

4.2 Objectives

To achieve this aim, the research will pursue the following objectives:

4.2.1 To develop a robust phishing detection model that leverages both traditional metadata and graph-based features extracted from email and network attributes.
4.2.2 To integrate explainable artificial intelligence (XAI) techniques and real-time cyber threat intelligence (CTI) within the detection pipeline.
4.2.3 To model authentication risk following phishing incidents through graph-based behavioral analysis and anomaly detection.

5. Research questions

The following research questions have been derived directly from the objectives of the study. They are designed to be precise, unambiguous, and sufficiently narrow to guide the investigation into developing and validating an explainable, phishing detection framework.

1. **To what extent does a metadata-driven, graph-based approach improve phishing detection accuracy compared to traditional machine learning models?**
2. **How does the combined integration of SHAP explanations and real-time CTI enrichment affect the interpretability, accuracy, and responsiveness of the phishing detection framework?**
3. **To what extent can post-phishing authentication behaviors such as anomalous login patterns and geolocation inconsistencies serve as reliable indicators of account compromise?**

6. Methodology

This section outlines the methodology that will be used to develop and evaluate the proposed explainable phishing detection and authentication risk modeling framework. The methodology follows a positivist, quantitative research approach aimed at designing, implementing, and empirically validating a lightweight, AI solution.

6.1 Research Design

Research Paradigm: This study adopts a positivist research paradigm, which is grounded in objectivity, reproducibility, and measurable outcomes. It is suitable for an Honours project because it emphasizes scientific rigor and supports empirical testing of hypotheses using observable and quantifiable data.

Approach: The project follows a deductive approach, testing hypotheses based on existing literature in phishing detection, graph-based modeling, and explainable machine learning.

Overview of Method Steps:

- The process will begin with the acquisition and preprocessing of labeled datasets. Confirmed phishing URLs will be sourced from PhishTank (2023) and validated using the VirusTotal API to reduce label noise. Benign samples will be extracted from the Enron Email Corpus, further

augmented with reputable domain lists such as the Alexa Top 1M to ensure a reliable control group.

- In the second stage, relevant metadata will be extracted from each email instance. This includes lexical features from URLs, sender-domain attributes, and structured email header information. These features will be used to construct a multipartite graph using NetworkX, linking entities such as senders, URLs, and domains. From this graph, structural properties including node degree, betweenness centrality, and clustering coefficients will be computed to capture topological indicators of malicious activity.
- Each sample will then be enriched with dynamic threat intelligence. Real-time indicators of compromise (IOCs) will be retrieved via the VirusTotal API, adding a contextual threat dimension to the feature space. This step integrates up-to-date risk signals, allowing the model to reason about evolving phishing patterns.
- A Random Forest classifier will be trained using the full feature set, which includes both engineered metadata and graph-derived metrics. To ensure interpretability, tree-SHAP (SHapley Additive exPlanations) will be employed to generate per-instance feature attributions. These explanations will enable transparency in the model's decision-making process, aligning outputs with domain-relevant reasoning.
- To assess the framework's capacity to model post-compromise behaviour, synthetic authentication logs will be generated. These logs will include timestamped login records, anomalous geolocation shifts, and device identifier changes based on behavioral baselines extracted from prior literature. The resulting behavioral features will be integrated into a secondary risk scoring module, simulating account compromise detection.
- The final phase will involve a comprehensive evaluation of the framework. Standard classification metrics including accuracy, precision, recall, F1-score, and ROC-AUC will be reported using 5-fold cross-validation. SHAP value distributions will be analysed to assess explanation clarity and domain alignment.

6.2 Research Method

The research will use experimental design as the core method, evaluating the performance of different model configurations on phishing datasets enriched with real-time threat data and behavioral features.

The framework includes:

- Feature Engineering using metadata and graph structures.
- Model Training using Scikit-learn's Random Forest implementation.
- Explainability Integration using SHAP.
- IOC enrichment through the VirusTotal API (VirusTotal, 2023).

6.2.1 Data Collection, Preparation, and Preprocessing

Datasets:

- PhishTank (2023): PhishTank is a free, public platform offering verified phishing URLs. It will serve as the main source of phishing samples. To improve reliability, all URLs will be validated through the VirusTotal API to remove stale or mislabeled entries. The dataset will be periodically refreshed to maintain relevance to current phishing tactics.
- Enron Email Corpus (Klimt & Yang, n.d.): The Enron email dataset (available via https://www.cs.cmu.edu/~enron/) contains real-world corporate emails, making it a suitable

source for benign examples. It will be filtered to remove internal spam and combined with a whitelist of trusted domains (e.g., Alexa Top 1M) to ensure a clean baseline.

- Custom Authentication Logs: Custom behavioral logs will be generated to model post-phishing anomalies. These will include login timestamps, IP geolocations, and device fingerprints, based on patterns observed in prior work such as (Wang et al., 2024). No public dataset will be used; instead, logs will be created synthetically to simulate compromised vs. normal behavior.

Preprocessing Steps

1. Metadata Extraction: Parse URLs, sender domains, email headers.
2. Graph Construction: Use NetworkX to build sender→URL→domain graphs, compute node degree, betweenness, clustering coefficient.
3. Feature Engineering: Combine tabular metadata (URL length, domain age) with graph metrics.
4. Normalization & Split: Scale features to [0,1], then split into 70% train / 30% test sets.

6.2.2 Algorithms and Models

- ***Random Forest Classifier***: The Random Forest (RF) algorithm will be trained on the feature set derived from phishing and benign datasets (e.g., lexical URL features, email metadata, graph-based metrics). This ensemble method constructs multiple decision trees and aggregates their votes for robust classification. RF is well-suited for tabular and high-dimensional feature spaces, making it ideal for structured metadata.

  It also supports feature importance extraction and integrates well with SHAP for explainability, directly supporting Objective 1 and 2 (accurate phishing detection and model transparency).The framework must not only detect phishing but do so in a way that is interpretable to analysts. RF + SHAP helps us detect phishing patterns while also exposing model reasoning, addressing the need for transparency identified in our problem statement.

- ***SHAP (SHapley Additive Explanations):*** SHAP will be used to generate per-sample feature attribution explanations from the Random Forest model. For each phishing or benign classification, SHAP will identify the features that most contributed to the decision (e.g., suspicious URL tokens or unusual domain names).

  By providing human-readable reasoning for each decision, SHAP supports explainability a critical requirement in cybersecurity. This helps analysts trust the output and verify that decisions align with known phishing heuristics (e.g., misspelled domains, spoofed senders). One key challenge is model explainability in dynamic environments. SHAP helps address this by making model predictions interpretable, directly supporting Objective 2 and contributing to solving the lack of transparency in phishing detection systems.

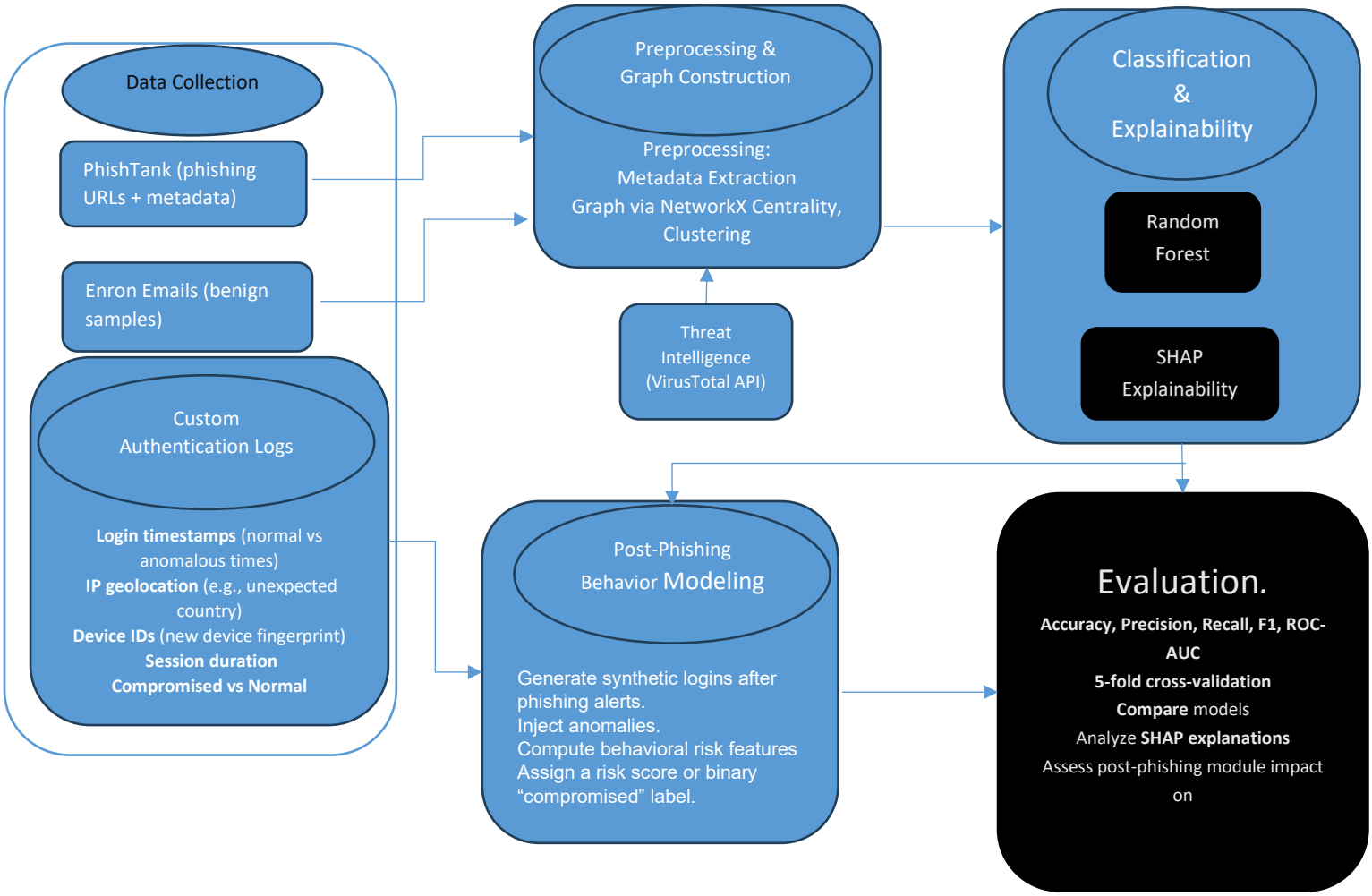
- ***NetworkX Graph Modeling***: Graphs G=(V,E)G = (V, E) are constructed where nodes are senders, URLs, and domains. NetworkX will be used to build multipartite graphs linking senders, domains, and URLs. Each email or URL becomes part of a connected graph, from which structural features like node degree, clustering coefficient, and betweenness centrality are extracted.Graph features can expose hidden relationships between phishing campaigns (e.g., shared infrastructure or coordinated domain use).

This enhances the feature set beyond what lexical or metadata alone can offer, supporting Objective 1 (improving detection performance with structural insights).Our study targets phishing that uses subtle domain and network-level deception. Graph analysis helps uncover these patterns, supporting detection of phishing at the infrastructure level a gap identified in the literature and our problem statement.

- ***VirusTotal API Integration***: Live URL/domain IOC validation is performed by querying VirusTotal's API to enrich graph nodes with dynamic threat scores (VirusTotal, 2023). For each URL or domain, VirusTotal will be queried in real-time to obtain dynamic threat scores (e.g., blacklisting status, malware association). This intelligence will be used to enrich features during both training and inference stages.

  Threat intelligence ensures the model remains context-aware and up to date. It also helps validate stale entries from datasets like PhishTank and augments detection with third-party insights.Phishing is a fast-evolving threat. Static models trained on outdated data underperform. By incorporating CTI (Cyber Threat Intelligence), the system gains adaptability and relevance, directly supporting Objective 2 and answering Research Question 3 on CTI integration.

Explainable Phishing Detection & Authentication Risk Modeling Pipeline Diagram



6.3 Data/Results Analysis

6.3.1 Model Performance Metrics

To assess the performance of the proposed phishing detection framework, we will employ a suite of widely accepted classification metrics: **Accuracy, Precision, Recall, F1-Score,** and **ROC-AUC**. These metrics are standard in cybersecurity machine learning tasks, particularly in phishing detection, and are chosen to reflect both predictive power and practical applicability in Security Operations Center (SOC) environments.

- **Accuracy** measures the overall proportion of correct predictions (both phishing and benign). It provides a general benchmark for model performance. However, in imbalanced datasets, common in phishing detection, the accuracy alone can be misleading, as a model can achieve high accuracy by simply predicting the majority class. Therefore, we report it alongside more discriminative metrics.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}}$$

- **Precision** reflects the proportion of true phishing predictions out of all emails classified as phishing. In a SOC setting, high precision is essential to minimize false positives, which can overwhelm analysts and reduce trust in the detection system. According to (Kavya & Sumathi, 2024) research, precision is a crucial measure in real-time phishing filters to prevent unnecessary alert fatigue. This is because high precision means fewer false positives, which translates to a more accurate and reliable system, preventing users from being overwhelmed by alerts that are not actual phishing attempts.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

- **Recall (Sensitivity)** measures the model's ability to detect actual phishing attempts. A high recall ensures that few malicious emails are missed, which is vital in high-risk environments where undetected threats can lead to breaches. (Lim et al., 2025) highlight recall as a crucial metric in phishing detection due to the high risk associated with failing to detect phishing attacks (false negatives). False negatives, which occur when a phishing email is incorrectly identified as legitimate, can lead to significant consequences, including financial loss, data breaches, and reputational damage.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

- **F1-Score** offers a **harmonic mean** between precision and recall, balancing the trade-off between missing attacks (FN) and triggering too many false alarms (FP). This makes it especially useful when evaluating model performance in imbalanced datasets. (Elkouay et al., 2024) and (Thakur et al., 2023) report F1-Score as a primary metric for phishing classifiers, as it synthesizes the most relevant aspects of model reliability.

$$\boldsymbol{F1} = \boldsymbol{2} \cdot \frac{\boldsymbol{Precision} \cdot \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

- **ROC-AUC(Receiver Operating Characteristic – Area Under Curve)-** measures the model's ability to distinguish between phishing and benign classes across all classification thresholds. AUC values close to 1.0 indicate excellent separability. This metric is especially valuable when comparing model performance across different configurations or feature sets.

Why These Metrics?

- These metrics are **standardized in phishing literature**, enabling comparison with prior studies such as (Remya et al., 2025) who used precision, recall, and F1 to evaluate their BGL-PhishNet framework.
- **Precision and Recall** directly map to real-world SOC objectives: reducing alert fatigue and minimizing missed attacks.
- **F1 and ROC-AUC** offer a balanced and threshold-independent evaluation, especially useful in comparing model variants across experiments.

6.3.2 Statistical Methods

k-Fold Cross-Validation (k=5): To evaluate model generalization, we apply stratified 5-fold cross-validation, ensuring that each fold maintains the same proportion of phishing and legitimate samples. This is essential for imbalanced datasets like phishing detection, as it mitigates the risk of biased metric estimates due to class imbalance (Pedregosa , 2011)

We selected 5 folds to strike a balance between computational efficiency and evaluation reliability, consistent with common practice in machine learning research (Hastie et al., 2009). For each fold, performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC will be calculated and reported as mean ± standard deviation across all folds.

To assess the statistical significance of observed differences between model variants (e.g., metadata-only vs. graph-enhanced), we will use paired t-tests on per-fold scores. In addition, feature importance will be assessed using SHAP value distributions. Global SHAP summaries will be used to identify consistently influential features, while local SHAP explanations will be examined for selected predictions to verify whether they align with known phishing indicators (e.g., “domain age = young”, “sender mismatch = yes”). If feasible, domain experts may be consulted to validate explanation relevance and clarity.

Finally, we will conduct a comparative analysis against baseline models that exclude graph-based features and/or explainability, to quantify the added value of each component in the proposed framework.

6.3.3 System Specifications

- Development Environment: Ubuntu 22.04, Python 3.10
- Libraries: Scikit-learn, SHAP, NetworkX, Boto3, Pandas, Matplotlib
- Hardware: 8-core CPU, 16GB RAM (for local testing)

7. Expected Outcomes

This research is expected to generate both theoretical and practical contributions to the domains of phishing detection, explainable artificial intelligence (XAI), and cyber threat intelligence (CTI) integration. The projected outcomes are outlined below.

7.1 Technical and Academic Contributions

7.1.1 An Explainable, Lightweight Phishing Detection Framework

The study aims to develop a prototype phishing detection system that leverages metadata link graphs, real-time CTI enrichment, and SHAP-based interpretability mechanisms. This

approach is designed to address a notable gap in current cybersecurity systems, where high accuracy often comes at the expense of transparency and operational readiness. The system is expected to enable more intuitive model outputs that are interpretable by human analysts.

7.1.2 A Novel Integration of CTI and Post-Phishing Risk Modeling

The research proposes a pipeline that integrates real-time indicators of compromise (IOCs), such as VirusTotal feeds, with behavioral anomalies following phishing attempts (e.g., unusual login patterns). This integration is expected to support a more dynamic and context-aware risk assessment process, offering a layered defense beyond traditional phishing detection.

7.1.3 Contribution to Explainable AI Literature in Cybersecurity

By applying SHAP-based explanation techniques within a phishing detection context, the project intends to empirically evaluate their impact on interpretability and decision-making. While prior work has explored XAI in malware and intrusion detection, this study seeks to extend its application to phishing and authentication modeling, contributing to emerging literature in explainable threat analytics (Gearhart, 2024; Prity et al., 2024; Mohale & Obagbuwa, 2025).

7.2 Practical and Societal Impact

7.2.1 Enhanced Decision-Making in Security Operations Centers (SOCs)

The proposed framework is expected to enhance SOC workflows by providing analysts with actionable, human-readable threat explanations. Studies such as (IBM, 2024) suggest that such explainability can significantly reduce response time. Although the effect may not be directly measured in this study, the system design prioritizes usability and interpretability for operational effectiveness.

7.2.2 Proactive Defense Against Evolving, AI-Generated Threats

With phishing attacks increasingly using automated and generative AI techniques, this research aims to contribute to more adaptive and proactive defense mechanisms. By avoiding reliance on static heuristics or signatures, the framework is positioned to help security teams respond effectively to fast-evolving attack strategies.

7.3 Potential for Future Expansion

The architecture developed in this study is intended to be extensible and adaptable. In the future, it may be applied to other cybersecurity domains such as:

- Insider threat detection
- Malware classification
- Fraud analytics
- Integration with SOAR (Security Orchestration, Automation, and Response) platforms
  Future work could also explore the use of advanced behavioral modeling, dynamic user risk scoring, and further automation of threat response pipelines.

8. Timeline and Budget

8.1 Timeline

The research is expected to be completed over a **6-month period**, with defined phases for planning, implementation, evaluation, and reporting. Below is a table outlining the timeline for the

main phases of the project.
Research Timeline and Objectives

Research Phase	Objectives	Deadline
1. Background research and literature review	- Meet with supervisor for scope discussion - Read recent literature on phishing, XAI, CTI, and graph modeling - Identify research gaps - Refine research questions - Develop conceptual and theoretical framework	18th March
2. Research design planning	- Design framework architecture (phishing detection + auth behavior + CTI + SHAP) - Finalize toolset (e.g., NetworkX, SHAP, VirusTotal API) - Define datasets and ethical considerations for synthetic or anonymized user behavior data	30th May
3. Data collection and preparation	- Collect phishing datasets and authentication logs - Perform CTI enrichment using VirusTotal - Construct metadata graphs and label samples - Clean and pre-process data for modeling	15th June
4. Data analysis and model development	- Train Random Forest model using hybrid features - Apply SHAP to interpret detection outputs - Analyze authentication behavior post-phishing - Evaluate performance (accuracy, precision, interpretability)	30rd June
5. Writing	- Draft methodology, experiments, results, and discussions - Meet with supervisor for review and feedback - Address gaps or refinement in experimental setup	22nd August
6. Revision	- Complete 2nd draft with all refinements - Get supervisor approval for final document - Proofread, format, and prepare visuals - Submit final version and supporting code - Optional: Deploy or simulate prototype (if scope allows)	30th September

8.2 Budget

The project will primarily rely on open-source tools.

9. Ethical Considerations

This research does not involve any direct experimentation on humans or animals, nor does it require handling of chemicals, biological agents, or potentially hazardous materials. However, several ethical and procedural aspects will still be carefully considered in line with academic and institutional guidelines.

9.1 Data Privacy and Access

9.1.1 The study will utilize publicly available datasets such as the PhishTank phishing dataset and the Enron Email Corpus. These datasets are open-source and anonymized to prevent the disclosure of any personal or sensitive information.

9.1.2 Synthetic authentication data will be generated to simulate post-phishing behavior (e.g., login anomalies). This data will not be linked to any real individuals or systems.

9.2 Data Handling, Storage, and Sharing

9.2.1 All datasets and results will be stored securely on password-protected, encrypted local storage and backed up using approved university systems (e.g., OneDrive for Business or Google Drive with institutional access).

9.2.2 Any shared data or code (e.g., on GitHub) will exclude sensitive data and be governed by an open-source license (e.g., MIT or Apache 2.0) where appropriate.

9.2.3 Real-time threat data queried from sources like VirusTotal will comply with their API usage and privacy policies. No personally identifiable information (PII) will be stored.

9.3 Access and Use of Third-Party Services

9.3.1 API keys used (such as VirusTotal Public API) will be used responsibly and only for academic purposes, with proper authentication and rate-limiting measures in place.

9.4 Academic Integrity and Plagiarism

9.4.1 The entire research process will adhere to the institution's ethical policies on academic integrity.

9.4.2 A Turnitin plagiarism report will be submitted with the final research proposal as required.

9.4.3 The similarity index will be maintained below the threshold specified in the School Plagiarism Policy to ensure originality and proper citation of sources.

9.4.4 All external materials, including published articles, codebases, and datasets, will be properly cited.

9.5 Ethical Clearance

Based on the current design, the research does not require formal ethical clearance, as it does not involve human subjects, surveys, interviews, or interactive experimentation. However, should future phases of the research involve human feedback (e.g., usability testing of SHAP-generated explanations by Security Operations Center (SOC) analysts), a formal amendment will be submitted to the Sol Plaatje University Research Ethics Committee for approval.

10. References

Al, M., & Al Shwali, J. (n.d.). *Evaluation of Explainable AI Techniques for Interpreting Machine Learning Models*.

Ali, A., Li, J., Chen, H., Bhatti, U. A., & Khan, A. (2023). Real-Time Spammers Detection Based on Metadata Features with Machine Learning. *Intelligent Automation and Soft Computing*, *38*(3), 241–258. https://doi.org/10.32604/iasc.2023.041645

Al-Sabbagh, A., Hamze, K., Khan, S., & Elkhodr, M. (2024). An Enhanced K-Means Clustering Algorithm for Phishing Attack Detections. *Electronics (Switzerland)*, *13*(18). https://doi.org/10.3390/electronics13183677

Alturkistani, H., & Chuprat, S. (2024). *Artificial Intelligence and Large Language Models in Advancing Cyber Threat Intelligence: A Systematic Literature Review*. https://doi.org/10.21203/rs.3.rs-5423193/v1

Bardakis, A. (2024). *A Holistic Examination of the Methodology and Applications of Cyber Threat Intelligence*.

Divakaran, D. M., & Oest, A. (2022). *Phishing Detection Leveraging Machine Learning and Deep Learning: A Review*. http://arxiv.org/abs/2205.07411

Elkouay, A., Najem, M., & and Madani, A. (2024). Graph-based phishing detection: URLGBM model driven by machine learning. *International Journal of Computers and Applications*, *46*(7), 481–495. https://doi.org/10.1080/1206212X.2024.2342710

Fereidouni, H., Hafid, A. S., Makrakis, D., & Baseri, Y. (2024). *F-RBA: A Federated Learning-based Framework for Risk-based Authentication*. http://arxiv.org/abs/2412.12324

FRANK XAVIER GEARHART. (2024). *A Study on the Effectiveness of Interpretable Machine Learning Explanations in Cybersecurity*.

Gartner. (2024). *Gartner Identifies the Top Cybersecurity Trends for 2024*. https://www.gartner.com/en/cybersecurity

Guo, W., Wang, Q., Yue, H., Sun, H., & Hu, R. Q. (2025a). *Efficient Phishing URL Detection Using Graph-based Machine Learning and Loopy Belief Propagation*. http://arxiv.org/abs/2501.06912

Guo, W., Wang, Q., Yue, H., Sun, H., & Hu, R. Q. (2025b). *Efficient Phishing URL Detection Using Graph-based Machine Learning and Loopy Belief Propagation*. http://arxiv.org/abs/2501.06912

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Citeseer.

Huisman, J. (n.d.). *2024 Phishing Attack Landscape and Benchmarking The data you need to know*.

IBM Corporation. (2024). *Cost of a Data Breach Report 2024*.

Kavya, S., & Sumathi, D. (2024). Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review*, *58*(2), 50. https://doi.org/10.1007/s10462-024-11055-z

Klimt, B., & Yang, Y. (n.d.). *Introducing the Enron Corpus*. http://www-2.cs.cmu.edu/~enron/.

Li, K., Yang, T., Zhou, M., Meng, J., Wang, S., Wu, Y., Tan, B., Song, H., Pan, L., Yu, F., Sheng, Z., & Tong, Y. (2024). SEFraud: Graph-based Self-Explainable Fraud Detection via Interpretative Mask Learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 5329–5338. https://doi.org/10.1145/3637528.3671534

Lim, B., Huerta, R., Sotelo, A., Quintela, A., & Kumar, P. (2025). *EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability*. http://arxiv.org/abs/2503.20796

Maureen Oluchukwuamaka Okafor. (2024). Deep learning in cybersecurity: Enhancing threat detection and response. *World Journal of Advanced Research and Reviews*, *24*(3), 1116–1132. https://doi.org/10.30574/wjarr.2024.24.3.3819

Mia, M., Derakhshan, D., & Pritom, M. M. A. (2025). Can Features for Phishing URL Detection Be Trusted Across Diverse Datasets? A Case Study with Explainable AI. *Proceedings of the 2024 11th International Conference on Networking, Systems and Security, NSysS 2024*, 137–145. https://doi.org/10.1145/3704522.3704532

Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. In *Frontiers in Artificial Intelligence* (Vol. 8). Frontiers Media SA. https://doi.org/10.3389/frai.2025.1526221

Pedregosa FABIANPEDREGOSA, F., Michel, V., Grisel OLIVIERGRISEL, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, andÉdouard, & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Prity, F. S., Islam, Md. S., Fahim, E. H., Hossain, Md. M., Bhuiyan, S. H., Islam, Md. A., & Raquib, M. (2024a). Machine learning-based cyber threat detection: an approach to malware detection and security with explainable AI insights. *Human-Intelligent Systems Integration*, *6*(1), 61–90. https://doi.org/10.1007/s42454-024-00055-7

Prity, F. S., Islam, Md. S., Fahim, E. H., Hossain, Md. M., Bhuiyan, S. H., Islam, Md. A., & Raquib, M. (2024b). Machine learning-based cyber threat detection: an approach to malware detection and security with explainable AI insights. *Human-Intelligent Systems Integration*, *6*(1), 61–90. https://doi.org/10.1007/s42454-024-00055-7

Rastogi, N., Dhanuka, D., Saxena, A., Mairal, P., & Nguyen, L. (2025). *Survey Perspective: The Role of Explainable AI in Threat Intelligence*. http://arxiv.org/abs/2503.02065

Remya, S., Pillai, M. J., Aparna, B. S., Subbareddy, S. R., & Cho, Y. Y. (2025). BGL-PhishNet: Phishing Website Detection Using Hybrid Model-BERT, GNN, and LightGBM. *IEEE Access*, *13*, 47552–47569. https://doi.org/10.1109/ACCESS.2025.3551542

SlashNext. (2024). *Prepare for 2025 2024 Phishing Intelligence Report*. https://www.slashnext.com

Thakur, K., Ali, M. L., Obaidat, M. A., & Kamruzzaman, A. (2023). A Systematic Review on Deep-Learning-Based Phishing Email Detection. In *Electronics (Switzerland)* (Vol. 12, Issue 21). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/electronics12214545

Wang, C., Tang, H., Zhu, H., Zheng, J., & Jiang, C. (2024). Behavioral authentication for security and safety. *Security and Safety*, *3*. https://doi.org/10.1051/sands/2024003

Zhao, J., Yang, C., Wu, D., Cao, Y., Liu, Y., Cui, X., & Liu, Q. (2023). Detecting compromised email accounts via login behavior characterization. *Cybersecurity*, *6*(1). https://doi.org/10.1186/s42400-023-00167-8