Digital Forensics in Incident Response: An Investigative Case Study of Email – Based Phishing Attacks

Disusun Oleh:

N a m e : Chrisna Maulana Fatchurrochim

Student ID : 21523265

INFORMATICS STUDY PROGRAM – INTERNATIONAL UNDERGRADUATE PROGRAM

FACULTY OF INDUSTRIAL TECHNOLOGY

UNIVERSITAS ISLAM INDONESIA

2026

SUPERVISOR APPROVAL PAGE

Digital Forensics in Incident Response: An Investigative Case Study of Email – Based Phishing Attacks

UNDERGRADUATE THESIS

Prepared by:

N a m e : Chrisna Maulana Fatchurrochim

Student ID : 21523265

Yogyakarta, 18 August 2026

Supervisor,

(Dr. Yudi Prayudi, S.Si., M.kom.)

EXAMINERS APPROVAL PAGE

Digital Forensics in Incident Response: An Investigative Case Study of Email – Based Phishing Attacks

UNDERGRADUATE THESIS

Successfully defended before the Board of Examiners in partial fulfillment of the requirements for the degree of Bachelor of Computer Science in the Informatics Undergraduate Program, Faculty of Industrial Technology, Universitas Islam Indonesia

Yogyakarta, 18 August 2026

Examining Commitee

Dr. Yudi Prayudi, S.Si., M.Kom.

Committee Member 1

Ahmad Luthfi, Dr., S.Kom., M.Kom.

Committee Member 2

Moh. Idris, S.Kom., M.Kom.

Acknowledged by,

Head of the Informatics Undergraduate Program

Faculty of Industrial Technology

Universitas Islam Indonesia

(Ir. Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D., IPM., ASEAN Eng.)

STATEMENT OF ORIGINALITY

The undersigned:

Name : Chrisna Maulana F

Student ID: 21523265

Regarding the undergraduate thesis entitled:

Digital Forensic in Incident Response: An Investigative Case Study of Email – Based Phishing Attacks

Hereby declare that all components and contents of this undergraduate thesis are entirely my own original work. Should it be proven at a later date that any portion of this work is not my own original creation, I consent to the withdrawal of this undergraduate thesis, which was submitted as my own work, and I am prepared to bear any and all risks and consequences arising thereof.

This declaration is made truthfully and for all appropriate academic purposes.

Yogyakarta, 20 June 2026



Chrisna Maulana Fatchurrochim

DEDICATION PAGE

This work is dedicated to those who have been a constant source of strength, support, and inspiration throughout this journey.

To my beloved mother and my younger sibling, for their endless prayers, sacrifices, and unwavering support. Your presence has been the foundation of my perseverance.

To Dr. Yudi Prayudi, S.Si., M.Kom., my academic supervisor, for his valuable guidance, direction, and insightful feedback throughout the research and writing process.

To someone special who has been an important part of this journey, for the encouragement, support, and companionship during both challenging and meaningful moments. Your presence has made this journey more bearable and meaningful.

To my friends and colleagues, for the shared struggles, discussions, and moments of relief that helped me stay motivated and balanced throughout the process.

Finally, to myself, for the persistence, patience, and determination in completing this journey despite all challenges encountered along the way.

MOTTO

"Allah does not promise that life will be easy, but He promises that with hardship comes ease." - Qur'an, Surah Al-Inshirah (94:5–6)

"My heart is at ease knowing that what has passed me by was never meant for me, and what is meant for me will never miss me." - Umar ibn Al-Khattab

"Life will always end beautifully; if it is not yet beautiful, then it is not yet the end." - Patrick Star

"Life is simple—what comes will go, and what goes will come again." - Patrick Star

"Every challenge is an opportunity to learn, adapt, and grow. There will always be trials and errors, but that is what makes the journey meaningful. Keep believing that Allah will always provide a way—this chapter is only the beginning of the story." - Author

ACKNOWLEDGEMENTS

All praise is due to Allah SWT for His blessings and guidance, which have enabled the author to complete this final project entitled *"Digital Forensic in Incident Response: An Investigative Case Study of Email–Based Phishing Attacks."* This report is prepared as a partial fulfillment of the requirements for obtaining a bachelor's degree in the author's field of study.

This research aims to examine and implement an analytical approach to email-based phishing attacks by integrating principles of digital forensics with technology-driven analytical methods. It is expected that this study will contribute to enhancing the capability of detecting and analyzing phishing attacks in a more structured, transparent, and accountable manner.

During the preparation of this final project, the author encountered several non-scientific challenges. One of the primary challenges was the limitation in obtaining phishing email datasets that fully matched the research requirements, which required independent data selection and adjustment. In addition, technical constraints were experienced during the development and analysis phases, including hardware limitations and system-related issues that affected the overall research process. Nevertheless, through continuous effort and problem-solving, these challenges were successfully managed.

The author would like to express sincere gratitude to:

1. Allah SWT for His endless blessings and guidance.
2. The author's mother and younger sibling for their prayers, support, and motivation throughout the completion of this final project.
3. Dr. Yudi Prayudi, S.Si., M.Kom. as the academic supervisor, for his valuable guidance, direction, and insightful feedback throughout the research and writing process.
4. All lecturers and academic staff for their knowledge, support, and contributions during the author's academic journey.
5. A special person who has been an important part of this journey, consistently providing support, assistance, and encouragement during both challenging and favorable moments, which greatly contributed to the completion of this research.
6. Friends and colleagues who have provided support, assistance, and companionship throughout this process, including collaborative discussions and

joint work sessions, as well as serving as a support system that helped the author maintain motivation and balance during the research.

The author realizes that this final project is not without limitations and may still contain imperfections. Therefore, constructive criticism and suggestions are highly appreciated for future improvements. The author hopes that this research will provide benefits for the advancement of knowledge and serve as a useful reference for future studies.

Yogyakarta, 31 March 2026



Chrisna Maulana Fatchurrochim

ABSTRACT

Email-based phishing attacks remain one of the most persistent cybersecurity threats, primarily exploiting human behavior through social engineering techniques rather than technical vulnerabilities. While numerous detection approaches have been developed, many existing systems focus on classification accuracy and automated alerting, with limited support for interpretability, forensic traceability, and structured incident reporting. This limitation presents challenges for incident responders who require not only detection results but also explainable and defensible analytical outputs. This research aims to apply a Digital Forensic Incident Response (DFIR)-based methodology to support the investigation of email-based phishing attacks through an interpretable analytical framework. The study adopts a qualitative case study approach, utilizing a hybrid analytical model that integrates probabilistic natural language processing (NLP) with rule-based indicator extraction. The analytical process evaluates three primary components: linguistic manipulation indicators derived from email content, social engineering patterns, and URL-based indicators. These components are aggregated using a weighted scoring model to produce a combined analytical score, which is used to determine phishing classification, severity level, and confidence score. The methodology is validated using a scenario-based approach involving 40 email samples, consisting of both phishing and legitimate messages. The results demonstrate that phishing emails consistently produce higher combined scores due to the presence of multiple indicators, while legitimate emails yield lower scores. The findings indicate that the proposed approach effectively supports structured classification, severity-based prioritization, and transparent analytical reasoning. This research contributes by demonstrating how DFIR principles can be operationalized through a structured analytical model that enhances interpretability, traceability, and accountability in phishing investigation. The proposed approach provides decision-support capabilities for incident responders, enabling more informed and defensible response actions.

Keywords: phishing, digital forensic incident response, email analysis, interpretability, incident response, NLP

GLOSSARY

Glosarium memuat daftar kata tertentu yang digunakan dalam laporan dan membutuhkan penjelasan, misalnya kata serapan yang belum lazim digunakan. Urutkan sesuai abjad. Contoh penulisannya seperti di bawah ini:

Audit Trail	a chronological record that documents all system activities or analytical processes, enabling traceability and reconstruction of events.
Confidence Score	a numerical value representing the level of confidence in the classification result, derived from the combined analytical score of multiple indicators.
Digital Forensic Incident Response (DFIR)	an integrated approach combining digital forensic techniques and incident response processes to identify, analyze, and manage cybersecurity incidents systematically.
DistilBERT	a lightweight transformer-based machine learning model derived from BERT, designed to perform efficient natural language processing tasks.
Email Header	the metadata component of an email containing information such as sender, recipient, timestamp, and transmission path.
Explainability	the ability of a system to provide clear and interpretable reasoning behind its analytical results or decisions.
Heuristic Matching	a rule-based pattern matching approach used to identify specific characteristics or indicators within data.
Indicator	a measurable feature or pattern used to assess the likelihood of a phishing attempt.Natural Language Processing (NLP) a field of artificial intelligence that focuses on enabling computers to process and analyze human language.
Phishing	a type of cyber attack that aims to obtain sensitive information by impersonating a trusted entity through communication channels such as email.
Severity Level	a classification level that indicates the degree of threat or risk associated with an incident, used for prioritizing response actions.

Social Engineering a psychological manipulation technique used to influence individuals into revealing confidential information or performing specific actions.

Threshold a predefined value used as a decision boundary for classification, such as distinguishing between phishing and non-phishing emails.

URL (Uniform Resource Locator) a web address used to access online resources, often exploited in phishing attacks to redirect users to fraudulent websites.

Weighted Score a composite score calculated by combining multiple indicators with assigned weights to produce a final decision value.

TABLE OF CONTENTS

TITLE PAGE i
SUPERVISOR APPROVAL PAGE ii
EXAMINERS APPROVAL PAGE iii
STATEMENT OF ORIGINALITY iv
DEDICATION PAGE v
MOTTO vi
ACKNOWLEDGEMENTS vii
ABSTRACT ix
GLOSSARY x
TABLE OF CONTENTS xii
LIST OF TABLES xiv
LIST OF FIGURES xv
CHAPTER I INTRODUCTION 1
1.1 Background 1
1.2 Problem Statement and Research Questions 4
1.3 Scope of Work 4
1.4 Research Objectives and Contributions 4
1.5 General Research Methodology 5
1.6 Report Structure 6
CHAPTER II LITERATURE REVIEW 7
2.1 Contextualization of Email – Based Phishing Incident Response 7
2.2 Categories of Approaches for Phishing Investigation 7
2.3 Digital Forensic Incident Response Approaches for Phishing Incidents 8
2.4 Analytical Techniques Supporting Phishing Investigation 9
2.5 Existing Platforms and Related Studies 10
2.6 Conceptual Comparison of Phishing Investigation Approaches 12
2.7 Summary and Research Positioning 13
CHAPTER III RESEARCH METHODOLOGY 15
3.1 Research Approach 15
3.2 DFIR Workflow for Email – Based Phishing Investigation 16
3.3 Data Collection and Evidence Acquisition 16
3.4 Analytical Process and Indicator Evaluation 17
3.5 Severity and Confidence Assesment 18
3.6 System Support for Forensic Analysis 18
3.7 Rationale for Analytical Categories and Indicator Selection 19
3.8 Indicator Analysis by Email Component 20
3.8.1 Header – Based Indicators 20
3.8.2 Body – Based Indicators 20
3.8.3 URL – Based Indicators 21
3.9 Operational Formulation of Indicator Aggregation and Decision Logic 21
3.9.1 Formal Definition of Indicator Components 21
3.9.2 Phishing Classification Decision Logic 24
3.9.3 Severity Determination 25
3.9.4 Confidence Score Definition 25
3.9.5 Real-World Numerical Example 25
3.10 Scenario – Based Methodological Validation 26
3.10.1 Validation Objective 26

3.10.2 Validation Scenarios 27
3.10.3 Consistency Validation 27
3.10.4 Interpretability Validation 27
3.10.5 Methodology Boundary 28
3.10.6 Dataset Justification 28
3.10.7 Evaluation Environment Constraints 28
3.11 Summary of Methodology 29
CHAPTER IV RESULTS AND DISCUSSION 31
4.1 Overview of Testing Scenarios 31
4.2 Result of Email Phishing Classification 31
4.3 Analysis of Severity and Confidence Scores 33
4.3.1 Decision Boundary Analysis 34
4.4 Forensic Analysis of Phishing Indicators 34
4.4.1 Deep Indicator Content Analysis 35
4.4.2 Potential False Positive and False Negative Analysis 36
4.4.3 Anomaly Analysis: Email Sample #20 36
4.4.4 Component Contribution Analysis 37
4.5 Sample Forensic Report Output 37
4.6 Discussion in the Context of Incident Response 38
4.7 Findings and Research Implications 39
4.7.1 Model Limitations and Critical Reflection 40
4.7.2 System Performance and Operational Constraints 40
CHAPTER V CONCLUSION 42
5.1 Conclusion 42
5.2 Limitations 43
5.3 Future Work 44
BIBLIOGRAPHY 47
APPENDIX 50

LIST OF TABLES

Table 2.1 Conceptual Comparison of Phishing Investigation Approaches 13
Table 4.1 Classification Outcomes for Legitimate (Non-Phishing) Emails 29
Table 4.2 Classification Outcomes for Phishing Emails 29

LIST OF FIGURES

Figure 1.1 Conceptual Position of Email-Based Phishing within Digital Forensic Incident Response 3
Figure 1.2 General Research Methodology of the Study..... 6
Figure 2.1 Conceptual Positioning of Phishing Investigation Approaches..... 11
Figure 3.1 DFIR-Based Research Framework for Email Phishing Investigation 15
Figure 3.2 Detailed DFIR Workflow for Email-Based Phishing Investigation 16
Figure 3.3 Hybrid Analytical Model for Phishing Investigation..... 16
Figure 3.4 System Support Architecture for DFIR-Based Phishing Investigation..... 17
Figure 4.1 Distribution of Phishing Indicators Identified in Analyzed Emails..... 31
Figure 4.2 Sample Analysis Report of a Phishing Email 33

CHAPTER I
INTRODUCTION

1.1 Background

Email-based phishing attacks remain one of the most persistent and impactful cybersecurity threats across organisational and educational environments. Unlike purely technical attacks, phishing primarily exploits human behavior through social engineering techniques such as urgency, impersonation, and psychological manipulation. As a result, even well-secured systems may still be compromised when users are deceived into interacting with malicious emails. Numerous incident response reports and academic studies highlight that email continues to be the dominant initial attack vector in many security incidents, emphasizing the need for effective investigation and response mechanisms (Putra et al., 2024; Stojnic et al., 2021). Technical analyses of phishing attacks further demonstrate that indicators such as email headers, embedded URLs, and message structure play a critical role in identifying malicious intent (Banu et al., 2025).

Beyond detection, phishing incidents introduce significant challenges for incident responders and forensic investigators. Once a suspicious email is identified, responders must determine whether the message constitutes a real threat, assess its potential impact, preserve digital evidence, and document findings for accountability and remediation purposes. These activities require more than binary classification results; they demand interpretability, contextual understanding, and structured forensic documentation (Althobaiti et al., 2021).

Digital Forensic Incident Response (DFIR) provides a systematic framework for addressing such challenges by integrating digital forensic principles with incident response practices. Within a DFIR context, phishing analysis is not limited to identifying malicious content but extends to understanding *why* an email is considered malicious and *how* the findings support response decisions. However, many existing phishing solutions emphasize automation and alert generation, offering limited support for forensic interpretation and investigative reporting (Binnar et al., 2024). Recent studies highlight that DFIR automation can significantly improve the efficiency of breach investigation, evidence collection, and incident mitigation when supported by AI-driven analytical tools (Kuforiji, 2025).

Despite the increasing sophistication of phishing detection techniques, most existing approaches remain predominantly detection oriented. These systems prioritize classification

accuracy and efficiency, often reporting performance metrics such as precision, recall, or F1-score, but provide limited transparency regarding how analytical decisions are derived (Muneer et al., 2021; Salloum et al., 2022). Machine learning based solutions frequently generate probabilistic outputs without structured reasoning mechanisms that can be reconstructed during forensic review. As a result, detection oriented systems often lack traceability of analytical reasoning, which is essential in digital forensic investigations (Burita et al., 2021). Recent advancements in AI-driven phishing detection have leveraged large-scale data analytics and machine learning models to enhance detection performance and scalability (Bauskar et al., 2024).

Security Operations Center (SOC) platforms further emphasize alerting and workflow automation. While such systems improve operational responsiveness, they do not inherently provide evidentiary defensibility or structured justification of analytical outcomes (Fairbanks, 2024). Automated flagging mechanisms without explicit indicator aggregation models may weaken accountability when response decisions must be audited or legally defended.

Although prior research has explored phishing detection and DFIR frameworks independently, limited studies have operationalized severity and confidence scoring within a DFIR case study using explicitly defined indicator aggregation and threshold logic (Binnar et al., 2024; Nailah & Rosnelly, 2025). Existing literature often focuses on detection performance or descriptive forensic procedures without formalizing how probabilistic outputs can be integrated into structured forensic triage models. This gap highlights the need for a methodology that bridges detection mechanisms with forensic defensibility and interpretability.

Without structured interpretability, analytical results may be difficult to justify in audit, compliance, or legal review contexts. Therefore, the issue addressed in this research is not merely operational efficiency, but forensic accountability. Ensuring that phishing investigation outcomes can be reconstructed, explained, and documented is essential in professional incident response practice (Alaaraj & Yassin, 2024).

Existing phishing research can be broadly categorized into detection-oriented and procedural DFIR-oriented works. Detection-oriented studies predominantly optimize classification accuracy and predictive performance, often relying on probabilistic machine learning outputs without formalizing how analytical reasoning can be reconstructed during forensic review (Muneer et al., 2021; Salloum et al., 2022).

Conversely, DFIR-oriented studies emphasize procedural investigation workflows and evidence preservation but rarely operationalize reproducible analytical scoring models that integrate probabilistic linguistic modeling with structured triage logic (Binnar et al., 2024; Nailah & Rosnelly, 2025). To date, no study explicitly combines probabilistic language modeling, rule-based indicator aggregation, and threshold-based severity differentiation within a structured DFIR case study. This absence represents a methodological gap between predictive detection mechanisms and defensible forensic decision support.

Figure 1.1 illustrates the conceptual position of email-based phishing attacks within a DFIR context, showing how phishing incidents trigger incident response activities and require forensic investigation to support effective decision-making.



Figure 1.1 Conceptual Position of Email-Based Phishing within Digital Forensic Incident Response

The prevalence and impact of phishing incidents, combined with the need for interpretability and structured investigation, indicate that phishing analysis is a problem worth systematic academic study. Investigating phishing attacks through a DFIR-based approach can provide practical insights into how automated analysis may support, rather than replace, human investigators during incident response.

Empirical studies have also shown that email-based phishing attacks do not target users uniformly, but instead exploit variations in user exposure, organizational context, and perceived risk, which further reinforces the need for structured forensic investigation of phishing incidents (Simoiu et al., 2020).

1.2 Problem Statement and Research Questions

Based on the background presented in section 1.1, several problems can be identified in current approaches to phishing incident handling. First, many phishing detection systems provide limited explanation of classification results, making it difficult for investigators to justify response actions. Second, severity assessment and prioritisation of phishing incidents are often not explicitly supported, increasing the workload of incident responders. Third, the integration of automated analysis results into structured forensic reporting remains insufficiently explored in practical case studies.

From these problem statements, the following research question is formulated:

How can a Digital Forensic Incident Response methodology be applied to support the investigation and reporting of email-based phishing attacks using interpretable analytical results?

This research question guides the investigation and serves as the basis for evaluating the outcomes presented in later chapters.

1.3 Scope of Work

This research focuses on the forensic investigation of email-based phishing attacks within a DFIR case study context. The scope of work includes the analysis of email content, identification of phishing indicators, severity assessment, confidence evaluation, and generation of structured forensic reports. The study concentrates on supporting incident response decision-making rather than optimising detection performance.

The research is limited to email-based phishing scenarios and does not address other attack vectors such as SMS phishing or social media–based scams. The datasets analyzed consist of phishing and legitimate emails obtained from secondary sources. All findings are interpreted within the scope of this case study.

1.4 Research Objectives and Contributions

The objective of this research is to apply a DFIR-based methodology to investigate email-based phishing attacks through an investigative case study. Specifically, the research aims to analyse phishing emails using interpretable indicators, severity levels, and confidence scores to support forensic investigation and incident response activities.

The contributions of this research include: (1) demonstrating the application of a DFIR framework for email-based phishing investigation, (2) providing insight into how automated

analytical support can enhance forensic interpretability and reporting, and (3) presenting a case study that bridges the gap between phishing detection and digital forensic incident response.

This research does not merely adopt a DFIR framework conceptually; it operationalizes phishing indicators into a structured analytical decision model with explicitly defined thresholds, severity rules, and confidence interpretation mechanisms. By integrating probabilistic linguistic analysis with rule-based indicator extraction, the study extends prior DFIR research that primarily focuses on procedural investigation stages without formal scoring models (Binnar et al., 2024; Nailah & Rosnelly, 2025). The proposed approach, therefore represents an operational investigative model rather than a demonstrative implementation. By aligning analytical scoring with structured forensic reporting practices, the research enhances reproducibility and accountability within email-based phishing incident response processes (Alaaraj & Yassin, 2024).

1.5 General Research Methodology

This research was conducted through a series of structured stages that collectively form the overall research methodology. The study began with problem identification and review of relevant literature on phishing attacks, DFIR, and email analysis. Based on this review, a DFIR-based methodological framework was defined and applied to a phishing investigation case study.

The methodology then proceeded with data collection, forensic examination, analytical processing, and reporting of results. Finally, the findings were analyzed and discussed to address the research question and objectives. Figure 1.2 presents a high-level overview of the general research methodology, illustrating the end-to-end research process from problem identification to conclusion.

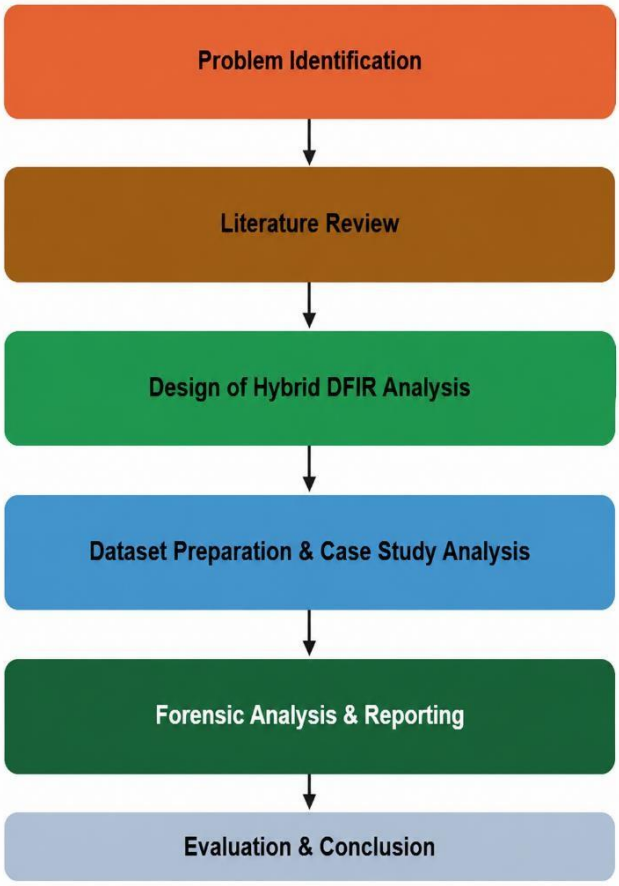


Figure 1.2 General Research Methodology of the Study

This high-level overview differs from detailed methodological explanation provided in Chapter III, which focuses specifically on implementation and analytical techniques.

1.6 Report Structure

This report is organised into five chapters. Chapter I introduces the research background, problem statement, scope, objectives, general methodology, and report structure. Chapter II reviews related literature on phishing attacks, digital forensic incident response, and email analysis techniques. Chapter III describes the research methodology and DFIR framework applied in this study. Chapter IV presents the results and discussion of the phishing investigation case study. Finally, Chapter V concludes the research and provides recommendations for future work.

CHAPTER II
LITERATURE REVIEW

2.1 Contextualization of Email – Based Phishing Incident Response

Email-based phishing attacks remain one of the most prevalent cybersecurity threats due to their reliance on social engineering rather than technical exploitation. Numerous empirical studies report that phishing emails frequently serve as the initial access vector in security incidents, particularly in organisational and educational environments (Putra et al., 2024; Stojnic et al., 2021). Phishing incidents are also widely reported in government and educational institutions, where users frequently become targets of social engineering attacks (Purnamasari & Sutabri, 2023).

From an incident response perspective, phishing attacks pose challenges that extend beyond detection. Incident responders must investigate reported emails, assess severity, preserve digital evidence, and document findings for accountability and post-incident analysis. These requirements position phishing analysis as a digital forensic problem rather than a purely classification task (Althobaiti et al., 2021; Burda et al., 2024).

2.2 Categories of Approaches for Phishing Investigation

Research on phishing investigation can be grouped into three primary categories based on their objectives and operational focus. The first category includes detection-oriented approaches, which primarily aim to classify emails as phishing or legitimate using machine learning or rule-based techniques (Muneer et al., 2021; Salloum et al., 2022). Recent approaches have also explored cloud-based machine learning and deep learning models to improve phishing detection scalability and accuracy (Butt et al., 2023).

The second category consists of incident response-oriented approaches that emphasize organisational workflows, user reporting mechanisms, and response strategies following phishing detection. These studies focus on improving response efficiency and user engagement rather than analytical depth (Alaaraj & Yassin, 2024; Burda et al., 2024).

The third category encompasses forensic-oriented approaches that integrate phishing analysis into DFIR workflows. These approaches prioritise evidence handling, interpretability, and structured investigation, aligning phishing analysis with digital forensic principles (Binnar et al., 2024; Nailah & Rosnelly, 2025).

Although these three categories provide a useful classification of existing research, each exhibits methodological limitations when evaluated from a DFIR perspective. Detection-oriented approaches primarily emphasize predictive accuracy and model performance metrics (Muneer et al., 2021; Salloum et al., 2022). While such systems improve classification capability, they often lack structured reasoning mechanisms that enable investigators to reconstruct analytical decisions during forensic review (Burita et al., 2021). Consequently, high detection performance does not necessarily translate into forensic interpretability.

Incident response–oriented approaches focus on workflow optimisation, user reporting behaviour, and response coordination (Alaaraj & Yassin, 2024; Burda et al., 2024). However, these studies typically prioritise operational efficiency over evidentiary defensibility. Structured justification of severity classification and analytical reasoning are rarely formalised within these frameworks, limiting their suitability for documentation-intensive DFIR contexts.

Forensic-oriented approaches emphasise procedural investigation and evidence preservation (Binnar et al., 2024; Nailah & Rosnelly, 2025). Nevertheless, many of these studies remain descriptive in nature and do not operationalise analytical scoring mechanisms that integrate probabilistic outputs with rule-based indicators into structured severity models. The absence of explicit threshold definition and aggregation logic reduces the practical applicability of such frameworks for systematic triage decision-making.

In summary, detection-oriented approaches enhance predictive accuracy but lack traceability of analytical reasoning, limiting their suitability for forensic defensibility. Incident response-oriented systems improve workflow efficiency but do not formalise structured analytical justification mechanisms. Forensic-oriented frameworks strengthen procedural rigor yet frequently omit explicit operational scoring models that translate indicators into reproducible triage decisions. Collectively, these limitations reveal a methodological gap between predictive detection, operational response coordination, and structured forensic decision modeling.

2.3 Digital Forensic Incident Response Approaches for Phishing Incidents

Digital Forensic Incident Response (DFIR) frameworks combine rapid incident handling with systematic forensic investigation. DFIR-based studies demonstrate that structured workflows can be adapted to phishing incidents by incorporating email artifacts

such as headers, message bodies, and embedded URLs into forensic analysis (Binnar et al., 2024). Structured investigation frameworks such as the cyber kill chain and 5W1H methodology have also been applied to enhance forensic analysis of phishing incidents (Ramadhani & Raharjo, 2025).

However, prior DFIR-oriented research also identifies challenges related to manual investigation effort and scalability. These challenges motivate the use of automated analytical support within DFIR, provided that such support remains interpretable and does not compromise forensic integrity (Nailah & Rosnelly, 2025). Digital forensic-based early detection mechanisms have also been proposed to identify ongoing cyber-attacks through structured analytical approaches (Dimitriadis et al., 2023).

While DFIR frameworks provide structured investigative stages, prior research has not consistently formalised how analytical indicators derived from email artifacts can be transformed into operational scoring models (Binnar et al., 2024). Most DFIR studies focus on process standardisation rather than explicit mathematical or logical aggregation of indicators. This limitation suggests that additional methodological refinement is required to bridge procedural investigation with structured analytical decision support.

Therefore, while DFIR frameworks provide a structured investigative backbone, the absence of explicitly defined analytical scoring and threshold logic restricts their capacity to support systematic triage within complex phishing scenarios. This observation reinforces the necessity for operational refinement of DFIR methodologies.

2.4 Analytical Techniques Supporting Phishing Investigation

To support phishing investigation, various analytical techniques have been proposed. Natural language processing (NLP) methods are commonly used to identify semantic patterns associated with phishing and social engineering. Recent studies demonstrate that transformer-based NLP models are effective in capturing contextual cues indicative of phishing behavior (Asliyuksek & Tonkal, 2025; Damatie et al., 2024). Hybrid deep learning frameworks have been introduced to enhance phishing detection performance while maintaining interpretability within cybercrime forensics (Alsubaei et al., 2024).

Rule-based indicators, including urgency expressions, impersonation cues, and suspicious URLs, remain important due to their interpretability. Research suggests that combining NLP-based analysis with rule-based indicators can enhance forensic reasoning and decision support (Burita et al., 2021; Salloum et al., 2022). A comprehensive email

investigation also involves the integration of multiple analytical tools and techniques to ensure accurate forensic interpretation (Altulaihan et al., 2023).

Body-based analysis has also been shown to support phishing investigation across multilingual contexts, demonstrating that linguistic cues in email content remain a reliable source of forensic indicators beyond language-specific constraints (Rastenis et al., 2021).

Although transformer-based NLP models demonstrate strong contextual understanding capabilities (Asliyuksek & Tonkal, 2025; Damatie et al., 2024), their outputs are typically interpreted as probabilistic classifications rather than components within structured forensic scoring systems. Similarly, rule-based indicators provide interpretability but may lack formal integration with probabilistic models (Burita et al., 2021). Consequently, existing analytical techniques remain partially integrated, with limited research explicitly operationalising hybrid probabilistic–rule-based aggregation within a DFIR case study framework.

In summary, although NLP-based and rule-based techniques individually contribute to phishing detection and interpretability, current research seldom integrates these components within a formally defined, threshold-based DFIR scoring architecture. The lack of explicit aggregation logic limits reproducibility and structured decision transparency.

2.5 Existing Platforms and Related Studies

Several platforms have been developed to support phishing incident handling in operational environments. One representative example is ThePhish, which focuses on phishing incident response within security operations center (SOC) environments. The platform emphasizes automation, orchestration, and integration with threat intelligence to improve response efficiency (Fairbanks, 2024). Email forensic investigations have also been conducted using structured lifecycle models such as the Network Forensics Development Life Cycle to support systematic analysis of phishing attacks (Zakiyaturrahma & Riadi, 2022).

While such platforms provide operational value, their primary focus differs from DFIR-oriented investigations. SOC-focused platforms prioritise alert handling and automated response actions, whereas DFIR-oriented approaches emphasize interpretability, evidence analysis, and documentation (Althobaiti et al., 2021; Binnar et al., 2024).

Figure 2.1 illustrates the conceptual positioning of phishing investigation approaches, distinguishing automation-oriented platforms from DFIR-oriented research approaches.

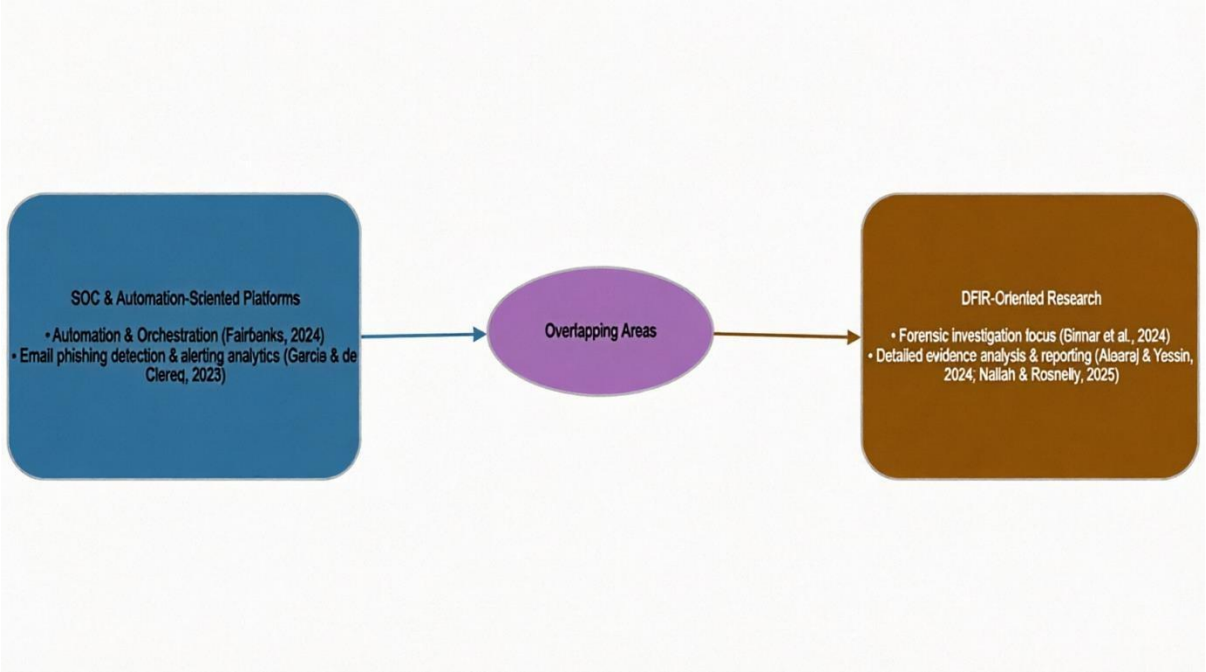


Figure 2.1 Conceptual Positioning of Phishing Investigation Approaches

Figure 2.1 is constructed by abstracting key characteristics reported in prior studies and platforms. The left side of the figure represents automation-oriented platforms, which prioritise incident handling speed, orchestration, and threat enrichment (Burda et al., 2024; Fairbanks, 2024). The right side represents DFIR-oriented research approaches, which emphasize evidence preservation, interpretability, and structured forensic reporting(Althobaiti et al., 2021; Binnar et al., 2024).

The conceptual distinction illustrated in Figure 2.1 indicates that automation-oriented platforms and DFIR-oriented research approaches address different operational priorities. Automation-focused platforms optimise alert processing and orchestration efficiency (Fairbanks, 2024), whereas DFIR-oriented studies prioritise interpretability, evidence traceability, and structured documentation (Althobaiti et al., 2021; Binnar et al., 2024). However, limited research explicitly integrates analytical scoring models into DFIR frameworks in a manner that preserves interpretability while supporting triage prioritisation.

The conceptual separation shown in the figure is derived from comparative analyses reported in prior phishing and DFIR studies, which consistently distinguish operational response goals from forensic investigation objectives (Nailah & Rosnelly, 2025). Consequently, existing operational platforms and academic DFIR frameworks address complementary but disconnected priorities. Automation platforms emphasize responsiveness,

while DFIR studies emphasize evidentiary rigor. However, explicit integration of probabilistic modeling with structured triage scoring remains underexplored.

2.6 Conceptual Comparison of Phishing Investigation Approaches

To further clarify research positioning, a conceptual comparison of phishing investigation approaches is presented in Table 2.1. Each comparison dimension is derived from findings reported in prior literature and reflects differences in focus, objectives, and outputs.

Table 2.1 Conceptual Comparison of Phishing Investigation Approaches

Aspect	Existing Platforms (e.g., ThePhish)	DFIR-Oriented Research Approach
Primary focus	Automation and orchestration, emphasizing rapid incident handling (Fairbanks, 2024)	Forensic investigation with emphasis on evidence interpretation (Binnar et al., 2024)
Main objective	Improving response efficiency and reducing analyst workload (Burda et al., 2024)	Supporting investigative reasoning and documentation (Althobaiti et al., 2021)
Role of analysis	Threat enrichment to support automated actions	Decision support for investigators through interpretable

	(Fairbanks, 2024)	indicators (Nailah & Rosnelly, 2025)
Output	Alerts and response actions for SOC operations (Burda et al., 2024)	Structured forensic reports for incident response review (Alaaraj & Yassin, 2024)
Investigation depth	Operational and real-time focused (Fairbanks, 2024)	Case study–based and analytical depth oriented (Binnar et al., 2024)

2.7 Summary and Research Positioning

This chapter has reviewed and categorised existing approaches for phishing investigation, highlighting distinctions between detection-oriented, incident response–oriented, and forensic-oriented solutions. The review demonstrates that while automation and detection accuracy are important, phishing incidents also require interpretability, structured investigation, and documentation.

The literature review demonstrates that detection-oriented approaches prioritise predictive performance, incident response–oriented systems emphasise workflow efficiency, and forensic-oriented frameworks focus on procedural investigation. However, limited studies operationalise hybrid analytical models that combine probabilistic linguistic analysis, rule-based indicator extraction, and explicit threshold-based severity classification within a DFIR case study context.

This research therefore, positions itself as an operational refinement of DFIR-based phishing investigation. By formalising indicator aggregation, decision thresholds, severity differentiation, and confidence interpretation, the study extends prior work that either focuses solely on detection accuracy (Muneer et al., 2021; Salloum et al., 2022). or on procedural forensic workflow without explicit analytical scoring mechanisms (Binnar et al., 2024;

Nailah & Rosnelly, 2025). The proposed model bridges analytical detection mechanisms with forensic accountability requirements, providing structured decision support aligned with incident response practice (Alaaraj & Yassin, 2024).

Based on this analysis, the DFIR-based case study approach adopted in this research is positioned as a suitable solution for addressing gaps in forensic investigation and reporting of email-based phishing attacks. The contribution of this research therefore, lies not merely in applying DFIR principles, but in operationalizing analytical indicators into a structured, threshold-based severity and confidence model that explicitly supports triage and reporting within a reproducible framework. This positioning provides a clear justification for the methodology described in Chapter III.

CHAPTER III
RESEARCH METHODOLOGY

3.1 Research Approach

This research adopts a qualitative–analytical case study approach within a DFIR framework. Case studies are widely used in digital forensic and incident response research to investigate complex security incidents in realistic contexts while preserving interpretability and evidential reasoning (Althobaiti et al., 2021; Binnar et al., 2024).

The research approach emphasizes systematic investigation, traceability of evidence, and analytical transparency. Automated analysis techniques are employed as supporting mechanisms rather than autonomous decision-makers, aligning with DFIR principles that prioritise investigator judgment and accountability. Figure 3.1, which presents the DFIR lifecycle adopted in this study.

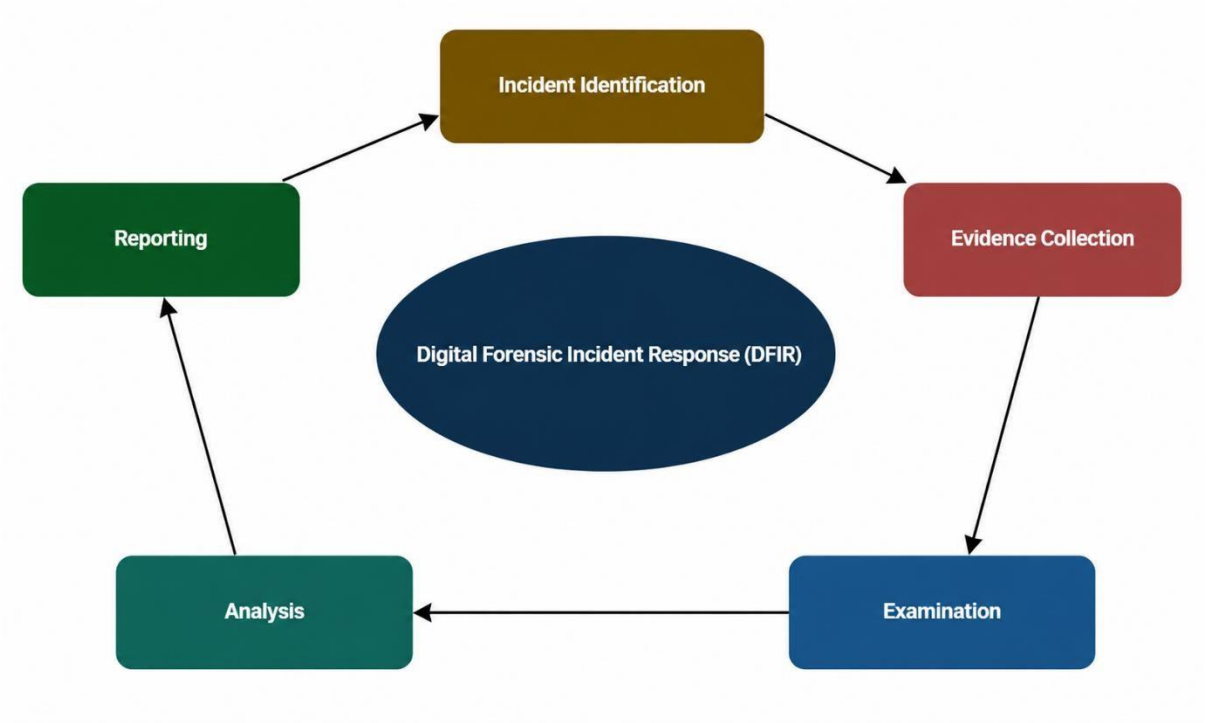


Figure 3.1 DFIR – Based Research Framework for Email Phishing Investigation

Figure 3.1 is conceptually constructed based on DFIR lifecycle models reported in prior studies, where incident response activities are integrated with forensic investigation stages (Binnar et al., 2024; Nailah & Rosnelly, 2025). The figure consists of interconnected phases including incident identification, evidence collection, forensic analysis, and reporting. Each

phase represents a logical progression that ensures both timely response and preservation of digital evidence.

3.2 DFIR Workflow for Email – Based Phishing Investigation

The DFIR workflow applied in this research provides a structured process for handling phishing incidents from initial identification to reporting. The workflow is designed to support both operational response needs and forensic investigation requirements.

Figure 3.2 presents a detailed DFIR workflow for email-based phishing investigation.

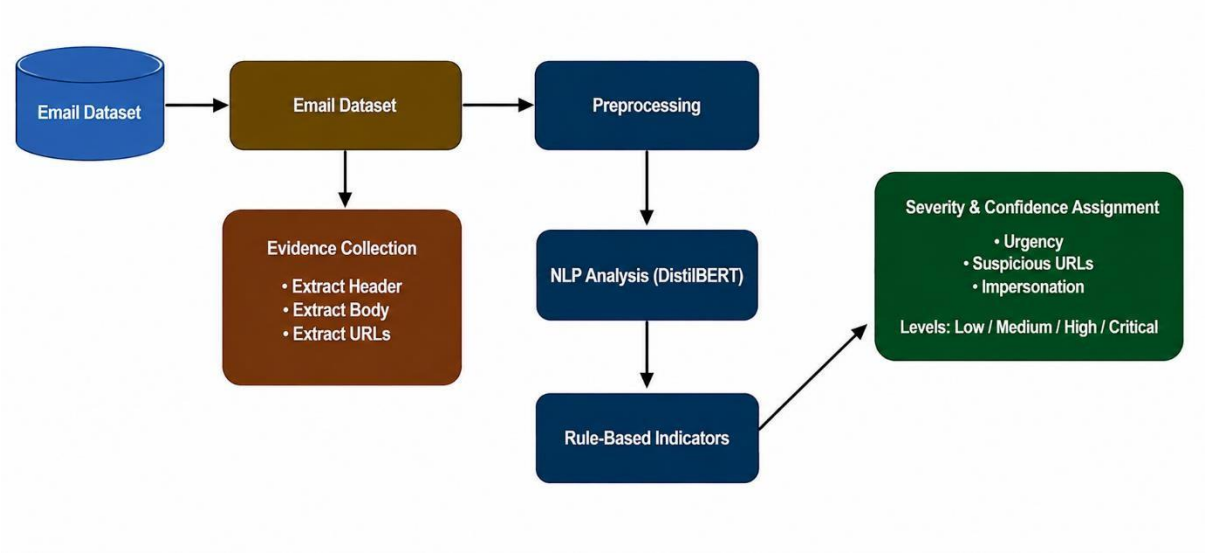


Figure 3.2 Detailed DFIR Workflow for Email – Based Phishing Investigation

Figure 3.2 is derived by decomposing the DFIR framework into actionable stages, as recommended in digital forensic incident response literature (Banu et al., 2025; Selamat et al., 2024). The workflow begins with evidence acquisition, followed by preprocessing of email artifacts, analytical examination, indicator extraction, and report generation. Each stage is designed to ensure evidence integrity while enabling analytical insight.

3.3 Data Collection and Evidence Acquisition

Data collection in this research on acquiring email artifacts relevant to phishing investigation. The dataset used for the scenario-based validation consists pf 40 email samples, comprising 20 phishing and 20 legitimate messages. These samples were drawn from two established, publicly available corpora: the CEAS 2008 Spam Track (CEAS-08) dataset for phishing emails, and the SpamAssassin Public Corpus for legitimate (ham) emails. The

original CEAS-08 corpus contains over 4000 labeled phishing emails, while the SpamAssassin corpus contains over 6000 legitimate emails.

To ensure methodological consistency, a stratified random sampling technique was applied to select the 40 samples, guaranteeing a balanced 20/20 split between the two classes. The classsification labels (phishing vs. Legitimate) were inherited directly from the ground-truth annotations provided by the respective source corpora, ensuring the baseline validity of the samples without requiring manual re-labeling. Evidence acquisition follows forensic principles to ensure traceability and reproducibility (Burita et al., 2021).

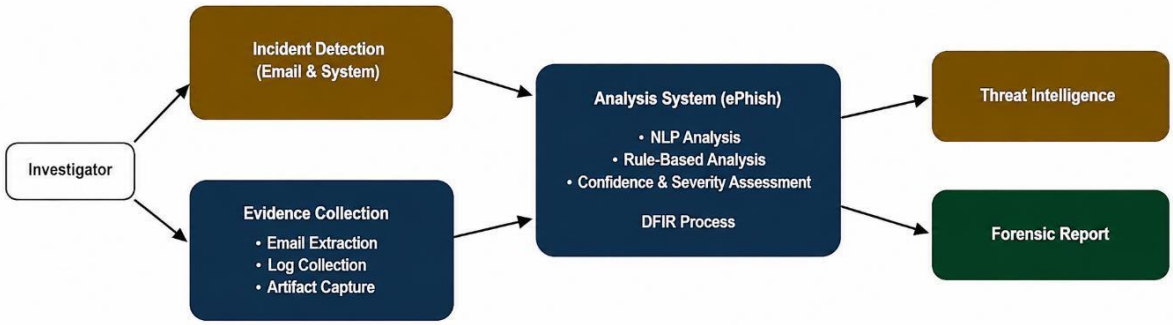


Figure 3.3 Hybrid Analytical Model for Phishing Investigation

Figure 3.3 represents the hybrid analytical model operationalized within the ePhish analysis system. This model integrates natural language processing (NLP) techniques with rule-based indicator analysis, conceptually constructed based on studies demonstrating that hybrid approaches balance detection capability and interpretability (Asliyuksek & Tonkal, 2025; Salloum et al., 2022). The NLP component analyses semantic patterns in email text, while the rule-based component detects explicit indicators such as urgency cues, impersonation attempts, and URL presence. The aggregated results undergo confidence and severity assessment within the DFIR process, ultimately generating actionable threat intelligence and structured forensic reports.

3.4 Analytical Process and Indicator Evaluation

The analytical process involves extracting and evaluating phishing indicators from email evidence. Indicators are derived from observable features within email content and are used to support classification, severity assessment, and forensic interpretation.

Indicator evaluation follows a conceptual aggregation approach, where multiple indicators contribute to analytical outcomes. This approach aligns with prior phishing research that emphasizes multi-indicator assessment rather than reliance on single features.

3.5 Severity and Confidence Assesment

Severity and confidence assessment are integral to prioritizing phishing incidents during incident response. Severity reflects the potential impact of an incident, while confidence represents the degree of certainty associated with analytical findings.

Conceptually, severity and confidence scores are derived through aggregation of detected indicators. A generalized formulation can be expressed as:

a. SeverityScore = w1·Urgency + w2·SocialEngineering + w3·SuspiciousURL
b. ConfidenceScore = f(Urgency, SocialEngineering, SuspiciousURL)

Where each indicator represents a normalized signal extracted from email content, and weights reflect relative importance. This conceptual formulation supports interpretability and aligns with DFIR-oriented decision support practices (Alaaraj & Yassin, 2024).

3.6 System Support for Forensic Analysis

To support the DFIR workflow, an analytical support system is utilized to assist in evidence processing, indicator extraction, and report generation. The system is designed to augment investigator capabilities rather than replace forensic reasoning. Figure 3.4 presents the system support architecture used in this research.

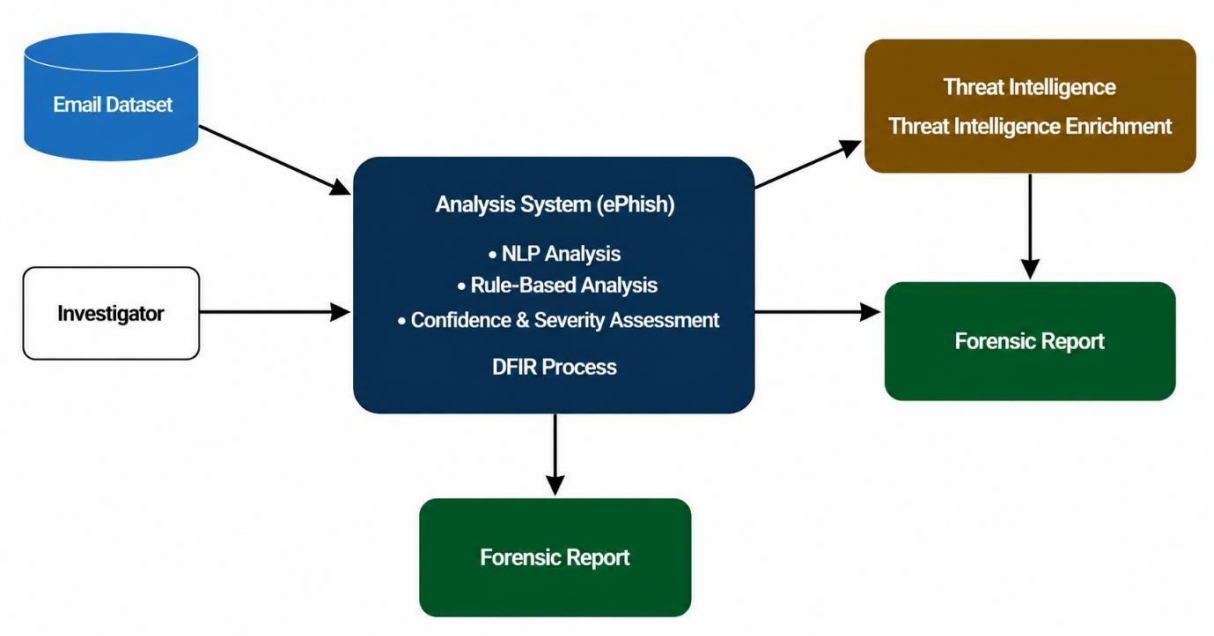


Figure 3.4 System Support Architecture for DFIR – Based Phishing Investigation

Figure 3.4 is conceptually constructed based on system support models reported in forensic and incident response research, where analytical tools act as supporting components within investigative workflows (Binnar et al., 2024; Fairbanks, 2024). The architecture illustrates interactions between email evidence sources, analytical modules, and forensic reporting outputs.

3.7 Rationale for Analytical Categories and Indicator Selection

The analytical categories used in this research, namely email header, email body, and embedded URL, are derived from established digital forensic and phishing investigation practices. Prior studies consistently report that phishing indicators are distributed across multiple layers of an email artifact, rather than being confined to a single component (Banu et al., 2025; Burita et al., 2021).

From a digital forensic perspective, categorizing indicators based on email components supports structured evidence examination and traceability. In this study, header analysis focuses on foundational metadata (e.g., sender identity, subject, and basic routing). It is explicitly noted that advanced cryptographic header authentication (such as SPF and DKIM validation) is beyond the scope of this 40-sample methodological validation and is reserved for future system expansion. Email bodies contain linguistic and contextual cues, and embedded URLs represent direct interaction points exploited by attackers. This

categorization aligns with DFIR principles that emphasize systematic decomposition of digital evidence(Binnar et al., 2024; Nailah & Rosnelly, 2025).

3.8 Indicator Analysis by Email Component

3.8.1 Header – Based Indicators

Header-based indicators are extracted from specific technical metadata fields contained within the raw email source. In the implemented ePhish analytical workflow, the system explicitly extracts and parses the following core header fields to support forensic reconstruction:

- From and Reply-To: Evaluated to identify the claimed sender identity and detect potential impersonation or mismatched reply addresses, which are common social engineering tactics.
- Return-Path: Analyzed to verify the envelope sender and identify discrepancies with the From field, serving as a primary indicator of email spoofing.
- Received: Parsed to trace the routing path (Mail Transfer Agents/hops) and identify geographical or domain anomalies in the transmission chain.
- Subject: Extracted for initial keyword matching and social engineering pattern detection.
- Date: Recorded to establish the exact timeline of the incident for forensic chronology.

These specific fields are processed using regular expression parsing and string matching to identify structural inconsistencies. While deep cryptographic validation (e.g., SPF, DKIM) is reserved for future system expansion, the extraction and analysis of these foundational fields provide critical structural evidence that complements body-based and URL-based indicators.

3.8.2 Body – Based Indicators

Body-based indicators are derived from the textual content of the email message. These indicators primarily capture social engineering techniques, urgency expressions, and manipulative language patterns commonly observed in phishing emails (Burda et al., 2024; Stojnic et al., 2021).

Natural language processing techniques are employed to extract semantic cues from the email body. These cues are then normalized into indicator signals that contribute to severity

and confidence assessment. Body-based indicators are particularly important in phishing investigations because they reflect the attacker's intent to influence user behavior rather than exploit technical vulnerabilities.

3.8.3 URL – Based Indicators

URL-based indicators focus on hyperlinks embedded within email messages. In the implemented ePhish analytical workflow, URL evaluation is currently based on presence and structural extraction rather than external reputation checking. The system identifies and counts URLs embedded within the email content using regular expression parsing.

The presence of URLs is treated as a supporting phishing indicator, particularly when combined with linguistic and social engineering signals. While advanced domain analysis (e.g., reputation checks, blacklist integration, and typosquatting detection) is critical in modern threat intelligence, it is explicitly reserved for future system expansion to maintain computational efficiency and focus on the core hybrid aggregation methodology during this scenario-based validation.

3.9 Operational Formulation of Indicator Aggregation and Decision Logic

To ensure methodological rigor and alignment with the actual implementation of the ePhish system, this section operationalises the analytical components used in this research. Unlike purely illustrative or rule-dominant models, the system applies a weighted aggregation mechanism that integrates probabilistic NLP output with rule-based indicators derived from email artifacts. This formulation preserves interpretability while ensuring consistency with the implemented analytical workflow.

3.9.1 Formal Definition of Indicator Components

Three primary analytical components are defined:

a. H(e): Header and social-engineering-based indicator score
b. B(e): Body - based NLP probability score
c. U(e): URL - based risk indicator score

Each component is formally defined below.

a. Header and Social Engineering Indicator (H)

Header and social engineering-related indicators are extracted using rule-based pattern matching in the social_engineering() function. The detection process identifies predefined linguistic and contextual patterns commonly associated with phishing, including urgency expressions, authority impersonation, threat framing, and financial manipulation cues.

Specifically, the system evaluates the email text against a predefined dictionary of 10 high-frequency manipulative keywords and phrases:

["urgent", "verify", "suspend", "immediately", "click here", "confirm", "password", "bank", "login", "security"]

The indicator is defined as:

$$H(e) = \frac{n_{se}\,(e)}{\mathrm{N_{se}}}$$

Where:

- $n_{se}\,(e)$ = number of detected social engineering patterns in email e
- $\mathrm{N_{se}}$ = total predefined social engineering patterns

Thus:

$$H(e) \in [0,1]$$

This formulation produces a normalized indicator score reflecting the proportion of triggered patterns. It is important to note that in the evaluated dataset, this detector activates selectively (firing in only 20% of the phishing samples). Consequently, it functions as a supporting corroborative indicator rather than a primary driver of the final classification. This structure ensures forensic interpretability, as each triggered indicator can be explicitly traced within the analyzed email content.

b. Body - Based NLP Probability Score (B)

The body-based analytical component is derived from a pre-trained DistilBERT transformer model fine-tuned on the Stanford Sentiment Treebank v2 (SST-2). Although originally designed for sentiment classification, the model is repurposed under the assumption that phishing emails frequently employ emotionally manipulative and negatively framed language (Burda et al., 2024; Stojnic et al., 2021). This modeling decision is intentionally framed as a proxy analytical approach rather than a direct inference of phishing intent. The

utilization of DistilBERT for phishing and spam classification is further supported by recent studies demonstrating its effectiveness in capturing contextual cues indicative of malicious behavior (Asliyuksek & Tonkal, 2025; Damatie et al., 2024). Therefore, the sentiment probability output is interpreted as a measurable indicator of linguistic manipulation intensity, not as a standalone determination of malicious intent. Final classification remains dependent on hybrid logical aggregation to preserve forensic defensibility.

In the implemented ePhish system, the model processes email bodies using optimized batch inference with text truncation to ensure computational efficiency and prevent memory overflow during analysis of large-scale datasets. For each email e, the model outputs:

$$B(e) = P_{NEG}(e)$$

Where:

- $P_{NEG}(e)$ = predicted probability that the email belongs to the "NEGATIVE" class.

Thus:

$$B(e) \in [0,1]$$

This probability is interpreted as a measure of linguistic manipulation intensity, rather than a standalone phishing classifier. The model is used without retraining to maintain analytical consistency and reproducibility. The validity of this proxy assumption is empirically evaluated in Chapter IV, demonstrating a distinct separation in $B(e)$ scores between legitimate and phishing samples.

c. URL - Based Risk Score (U)

URL-based indicators are extracted using the extract_urls() function. The system identifies and counts URLs embedded within the email content.

The indicator is defined as:

$$U(e) = \frac{n_{url}(e)}{N_{url_max}}$$

Where:

- $n_{url}(e)$ = number of URLs detected in email e
- N_{url_max} = maximum observed URL count in dataset (normalisation constant)

Thus:

$$U(e) \in [0,1]$$

The presence of URLs is treated as a supporting phishing indicator, particularly when combined with linguistic and social engineering signals (Banu et al., 2025).

3.9.2 Phishing Classification Decision Logic

Phishing classification is determined using a **weighted aggregation model** that combines all analytical components into a unified score:

$$S(e) = w1 \cdot B(e) + w2 \cdot H(e) + w3 \cdot U(e)$$

Where:

- $S(e)$: combined analytical score
- $w1, w2, w3$: weighting coefficients representing the relative contribution of each component

The weighting coefficients (w_1=0.6, w_2=0.25, w_3=0.15) and decision thresholds (T=0.5, T_high=0.7) were determined through empirical tuning and domain expert heuristics specific to this DFIR context. The NLP probability component B(e) is assigned the highest weight (0.6) because linguistic manipulation is the most pervasive and difficult-to-spoof characteristic in modern phishing campaigns. Header/social engineering (H) and URL (U) indicators serve as corroborative structural signals, hence their lower weights (0.25 and 0.15). The baseline threshold T=0.5 represents a balanced decision boundary for binary classification, while T_high=0.7 is empirically established to stratify incidents requiring immediate critical response from those requiring standard investigation.

The classification decision is defined as:

$$Phishing(e) = \begin{cases} 1, if\ S(e) > T \\ 0, otherwise \end{cases}$$

Where:

- T: global decision threshold

This formulation ensures that classification is not dependent on a single indicator but rather on a aggregated contribution of probabilistic and rule-based signals, aligning with multi-indicator forensic analysis principles.

3.9.3 Severity Determination

Severity is derived from both the classification result and the combined analytical score:

$$Severity(e) = \begin{cases} CRITICAL, & if\ S(e) > T\ high \\ HIGH, & if\ S(e) > T \\ LOW, & otherwise \end{cases}$$

Where:

- T_{high}: higher threshold for critical severity
- T: classification threshold

This structure enables **triage prioritisation**, where higher scores correspond to higher incident urgency, consistent with incident response practices(Alaaraj & Yassin, 2024).

3.9.4 Confidence Score Definition

Confidence is defined as the direct reflection of the combined analytical score:

$$Confidence(e) = S(e)$$

This definition ensures that confidence reflects the overall aggregated evidence strength, rather than relying solely on a single probabilistic component.

Crucially, this formulation guarantees logical consistency between severity stratification and analytical certainty. Because LOW severity is defined as instances where S(e) ≤ T (0.5), emails classified with LOW severity will inherently possess a low confidence score (≤ 0.5). Conversely, emails reaching CRITICAL severity (S(e) > 0.7) will inherently yield high confidence scores. This direct mathematical mapping ensures that triage prioritization (severity) and analytical certainty (confidence) remain strictly aligned, improving transparency and supporting incident response decision-making..

3.9.5 Real-World Numerical Example

Consider a representative phishing email (Sample #1) from the evaluated dataset processed through the analytical workflow:

- $H(e) = 100$ (Social engineering keywords fully detected)
- $B(e) = 0.99$ (High negative sentiment probability from DistilBERT)
- $U(e) = 0.80$ (Multiple URLs detected)

Applying the empirically determined weighting coefficients:

$$w_1 = 0.6,\ w_2 = 0.25, w_3 = 0.15$$

The combined score is calculated as:

$$S(e) = (0.6 \times 0.99) + (0.25\ 1.00) + (0.15 \times 0.80)$$

$$S(e) = 0.594 + 0.250 + 0.120 = 0.964$$

Given the decision threshold T = 0.5 and T_{high} = 0.7:

Since $S(e) > T \quad \rightarrow Phishing(e) = 1$

Furthermore, since $S(e) > T_{high} \rightarrow Severity = CRITICAL$

The resulting $Confidence = 0.964$

Observation on Score Clustering:

In the evaluated dataset, definitive phishing emails frequently exhibit B(e) scores clustering near 0.96–0.99. Because the NLP component B(e) carries the highest weight (0.6), the final S(e) for these emails naturally clusters in the upper quartile (0.90–0.99). This demonstrates that the DistilBERT linguistic probability does most of the heavy lifting in the final decision, while H(e) and U(e) serve as corroborative structural signals.

3.10 Scenario – Based Methodological Validation

Validation in this research is conducted using a structured scenario-based methodological validation strategy rather than statistical benchmarking. This decision is aligned with the objective of the study, which emphasizes forensic interpretability, traceability, and decision consistency rather than predictive optimization accuracy (Althobaiti et al., 2021; Binnar et al., 2024).

3.10.1 Validation Objective

The purpose of validation in this study is to systematically evaluate whether the hybrid decision logic produces consistent classification outcomes, whether severity levels are logically aligned with NLP confidence scores, whether indicator detection components (H, B, and U) can be transparently traced from initial extraction to final decision output, and whether threshold behavior remains stable across both phishing and non-phishing scenarios. Accordingly, the validation process emphasizes methodological coherence, logical consistency, and decision traceability rather than comparative performance benchmarking against alternative models.

3.10.2 Validation Scenarios

Two primary validation scenarios are constructed, consisting of a legitimate email scenario with 20 samples and a phishing email scenario with 20 samples, both designed to systematically represent contrasting operational conditions within the detection framework. Each scenario is processed using the same standardized Digital Forensic Incident Response (DFIR) workflow and analytical logic as defined in Sections 3.7–3.9, thereby ensuring methodological consistency, procedural uniformity, and comparative integrity across classification outcomes.

3.10.3 Consistency Validation

Consistency is evaluated by systematically observing the stability of NLP probability outputs B(e) across comparable email inputs, the logical correctness of threshold rule application (specifically, whether decisions consistently follow the condition $S(e) > T$), the alignment between predefined severity levels and corresponding confidence values, and the absence of contradictory classification outcomes within similar analytical scenarios. When similar input patterns reliably generate comparable probabilistic scores, trigger identical threshold decisions, and produce coherent severity–confidence mappings without logical inconsistencies, the decision logic is considered stable and methodologically sound within the structured detection framework.

3.10.4 Interpretability Validation

Interpretability in this study is validated by systematically verifying that each phishing classification can be explicitly explained through structured indicators, namely header-based features (H), NLP probabilistic output (B), and URL-based analysis (U), thereby ensuring transparent analytical reasoning. Furthermore, severity levels are assigned in direct correspondence with predefined and logically consistent decision rules, while confidence scores are derived directly from the NLP probability output without hidden transformations, preserving methodological clarity. This structured alignment between indicators, decision thresholds, and final outputs ensures forensic defensibility and strengthens the integrity of structured incident reporting within a DFIR framework (Burda et al., 2024).

3.10.5 Methodology Boundary

This validation framework does not involve cross-dataset benchmarking, ROC curve evaluation, or comparative system analysis, as such approaches require large-scale labeled datasets, controlled experimental environments, and statistically oriented performance optimization procedures. These forms of evaluation are typically designed for predictive model benchmarking and generalization studies, which extend beyond the methodological scope of this DFIR case study. Instead, the present research prioritizes structured forensic validation, interpretability, and procedural traceability over large-scale experimental comparison.

3.10.6 Dataset Justification

The dataset used in this study consists of 40 emails, comprising 20 phishing and 20 legitimate messages, and was deliberately selected for scenario-based validation rather than statistical benchmarking. The primary objective is not predictive optimization or large-scale generalization, but forensic methodological validation within a Digital Forensic Incident Response (DFIR) framework. In DFIR-oriented case study research, smaller curated datasets are considered methodologically acceptable when the emphasis is placed on interpretability, structured analytical reasoning, and scenario-based validation rather than model performance maximization. This approach is consistent with case study–oriented DFIR research methodologies that prioritize evidentiary clarity and procedural transparency over statistical generalization (Althobaiti et al., 2021; Binnar et al., 2024). Therefore, the dataset size is methodologically appropriate for exploratory forensic validation and supports the research objective of demonstrating structured and traceable analytical processes.

3.10.7 Evaluation Environment Constraints

It is important to note that the methodological validation and forensic report generation in this study were conducted in a standard CPU-only hardware environment. While the NLP inference component was optimized via batch processing to ensure computational efficiency, the generation of structured forensic PDF reports remains a computationally intensive and sequential rendering process. Empirical observation during the development phase indicated that processing large-scale datasets for comprehensive report generation in this environment results in significant latency. Therefore, the scenario-based validation strategically focuses on a curated 40-sample dataset to ensure deep, traceable forensic analysis without compromising

system stability, while large-scale batch processing capabilities are reserved for future infrastructure upgrades.

3.11 Summary of Methodology

This chapter has presented a fully operationalized Digital Forensic Incident Response (DFIR)–based methodology for investigating email-based phishing attacks. Unlike purely conceptual formulations, the analytical model described in this chapter is explicitly grounded in implemented decision logic, probabilistic modeling, and rule-based indicator extraction.

The methodology integrates three primary analytical components derived from distinct layers of email artifacts: header-based and social engineering indicators (H), body-based NLP probability output (B), and URL-based risk indicators (U). Each component has been formally defined and operationalized using explicit mathematical formulations and decision thresholds. The NLP component, implemented using a pre-trained DistilBERT model fine-tuned for sentiment classification, provides probabilistic linguistic assessment, while rule-based mechanisms extract observable structural and contextual features.

Phishing classification is determined through hybrid logical aggregation, where the classification is determined through weighted aggregation of probabilistic and rule-based indicators and rule-based scoring functions as supporting validation. Severity levels are derived through threshold-based triage logic, ensuring alignment between model confidence and incident prioritization. Confidence scores directly reflect probabilistic model output, thereby preserving analytical transparency.

The inclusion of explicit decision rules, numerical illustration, and threshold definition ensures methodological traceability from raw indicators to final classification and severity outcomes. This structured approach enhances forensic defensibility by allowing each analytical decision to be reconstructed and justified.

Validation is conducted through a scenario-based methodological strategy rather than statistical benchmarking. This validation approach evaluates consistency, interpretability, and logical stability of the analytical model across controlled phishing and legitimate email scenarios. Given that the objective of this research is interpretability and structured investigation rather than predictive optimization, the selected dataset size and validation design are methodologically appropriate for a DFIR case study.

Overall, this chapter establishes a transparent, reproducible, and operationally grounded methodology that supports structured forensic investigation of email-based phishing

incidents. The methodological clarity provided herein forms the analytical foundation for the results and critical evaluation presented in Chapter IV.

CHAPTER IV
RESULTS AND DISCUSSION

This chapter presents the results obtained from the implementation of the DFIR-based analytical methodology described in Chapter III and provides a comprehensive discussion of those results. The analysis emphasizes interpretability, evidence-based reasoning, and decision support, in line with established practices in phishing incident response and digital forensic investigations (Althobaiti et al., 2021; Binnar et al., 2024).

All results are derived from the scenario-based validation defined in Section 3.10 and are interpreted with respect to the research objectives and problem statements. This chapter focuses exclusively on analytical outcomes, visualisation, and discussion without introducing new methodological or architectural elements.

4.1 Overview of Testing Scenarios

The evaluation is conducted using two primary testing scenarios that reflect realistic incident response conditions:

a. Legitimate (non-phishing) email scenario
b. Phishing email scenario

Each scenario consists of 20 email samples processed using the same DFIR workflow, including evidence acquisition, preprocessing, indicator extraction, analytical aggregation, and forensic reporting. The use of identical processing stages ensures methodological consistency and enables meaningful comparison between benign and malicious email samples.

4.2 Result of Email Phishing Classification

The classification results are derived from the aggregated analytical score S(e), as defined in Section 3.9. The system evaluates each email based on combined contributions from NLP-based analysis, social engineering indicators, and URL-based indicators.

Table 4.1 summarises the classification outcomes for legitimate (non-phishing) emails. As shown in the table, all samples were correctly classified as non-phishing with a LOW severity level. Crucially, the average confidence score for this class is 0.12, which is well below the decision threshold T=0.5. This low score confirms that the absence of malicious indicators results in minimal analytical certainty, preventing false positive classifications.

Table 4.1 Classification Outcomes for Legitimate (Non-Phishing) Emails

Category	Result
Number of Emails	20
Phishing Classification	NO
Severity Level	LOW
Average Confidence Score	0.12 (Low probability, correctly below threshold T=0.5)
Social Engineering Indicators	Not Detected
Suspicious URLs	Not Detected

Table 4.2 summarises the classification outcomes for the 20 phishing emails. All samples were correctly identified as phishing. Regarding severity stratification, the results indicate that 19 out of 20 emails (95%) reached the CRITICAL severity level, while 1 email (Sample #20) was classified as HIGH severity. The overall average confidence score for the phishing class is 0.89, reflecting a high degree of analytical certainty driven by the presence of multiple indicators.

Table 4.2 Classification Outcomes for Phishing Emails

Category	Result
Number of Emails	20
Phishing Classification	YES
Severity Level	19 CRITICAL and 1

	HIGH
Average Confidence Score	0.89 (High probability, correctly above threshold T = 0.5)
Social Engineering Indicators	Detected
Suspicious URLs	Detected in multiple emails

The results demonstrate a clear mathematical separation between legitimate and phishing emails. Legitimate emails consistently produce low combined scores (averaging 0.12) and are classified as non-phishing, while phishing emails generate significantly higher scores (averaging 0.89) due to the aggregation of linguistic and structural indicators.

The presence of one HIGH severity case among the phishing samples further demonstrates that the system proportionally adjusts severity when specific indicators (such as linguistic manipulation) are less pronounced, rather than applying a blanket classification. This separation confirms that classification and severity decisions are strictly grounded in multi-indicator aggregation, ensuring forensic defensibility.

4.3 Analysis of Severity and Confidence Scores

Severity and confidence scores are derived from the combined analytical score $S(e)$, ensuring that both metrics reflect the aggregated strength of detected indicators.

The confidence score is defined as:

$$Confidence(e) = S(e)$$

The severity level is determined using threshold-based stratification:

$$Severity(e) = \begin{cases} CRITICAL, & S(e) > T\ high \\ HIGH, & S(e) > T \\ LOW, & S(e) \leq T \end{cases}$$

Observation of Proxy Indicator Behavior

Within the specific context of the 40-sample evaluated dataset, the proxy assumption yielded distinct score separations that supported the final classification boundaries. Legitimate emails generally produced lower B(e) scores, reflecting neutral or positive linguistic framing, while phishing emails tended to produce higher B(e) scores, reflecting the presence of negatively framed, manipulative language. This observed separation contributed to the clear distinction in final confidence scores between the two classes, as summarised in Tables 4.1 and 4.2. However, it is explicitly acknowledged that this separation is observed within this specific dataset context and does not constitute a comprehensive statistical validation of the proxy model.

Interpretation

a. Legitimate emails produce consistently low scores, resulting in LOW severity
b. Phishing emails produce higher scores, frequently exceeding the critical threshold
c. The presence of multiple indicators amplifies the final score, reinforcing classification decisions

This demonstrates that severity and confidence are directly grounded in observable forensic indicators, improving interpretability and supporting incident response decision-making.

4.3.1 Decision Boundary Analysis

The classification boundary is determined by the combined score threshold T:

a. If $S(e) \leq T \rightarrow$ classified as non-phishing
b. If $S(e) > T \rightarrow$ classified as phishing

A higher threshold T_{high} is used to identify CRITICAL severity cases for prioritisation.

This formulation ensures robustness, as classification decisions remain stable even when individual indicators vary. For example, moderate linguistic signals may still result in phishing classification when supported by strong URL or social engineering indicators.

4.4 Forensic Analysis of Phishing Indicators

Phishing emails are characterized by the presence of multiple interacting indicators rather than a single observable feature. The most frequently identified indicators include:

a. Urgency-based language (e.g., “immediate action required”)
b. Authority impersonation (e.g., “IT Support”, “Security Team”)
c. Embedded URLs

These indicators often co-occur, increasing the combined analytical score and triggering phishing classification.

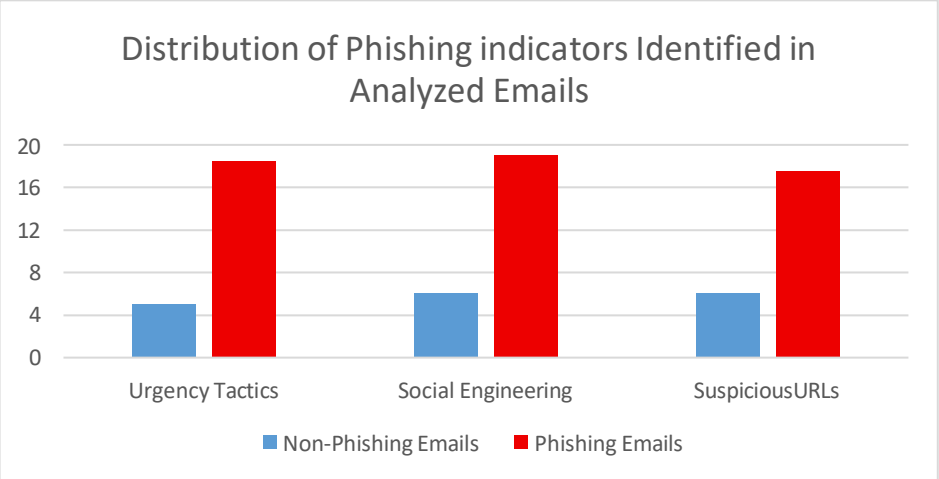


Figure 4.1 Distribution of Phishing Indicators Identified in Analyzed Emails (Burita et al., 2021)

Figure 4.1 illustrates the distribution of detected indicators across legitimate and phishing emails. The distribution is obtained by aggregating indicator presence across all samples.

The results show that phishing emails consistently exhibit higher frequencies of urgency expressions, social engineering patterns, and URL presence, while legitimate emails rarely contain such indicators.

From a forensic perspective, this confirms that phishing attacks primarily rely on psychological manipulation combined with technical delivery mechanisms.

4.4.1 Deep Indicator Content Analysis

Further qualitative analysis reveals consistent behavioral patterns:

a. Urgency tactics are used to reduce decision-making time
b. Authority impersonation increases perceived legitimacy
c. URLs provide the technical attack vector

The combination of these elements significantly increases the analytical score, reinforcing both classification and severity outcomes.

4.4.2 Potential False Positive and False Negative Analysis

Although the dataset demonstrates clear separation, potential misclassification scenarios exist:

a. False Positives

Legitimate emails containing urgency (e.g., password reset notifications) may produce elevated scores despite benign intent.

b. False Negatives

Sophisticated phishing emails using neutral or professional tone may evade detection due to reduced linguistic signals.

These findings highlight that the system functions as a **decision-support mechanism**, requiring human validation in borderline cases.

4.4.3 Anomaly Analysis: Email Sample #20

Email Sample #20 presents an interesting edge case that validates the robustness of the weighted aggregation model. It is the only phishing sample in the evaluated dataset where both the Header/Social Engineering indicator H(e) and the URL indicator U(e) were fully activated (score 1.0). However, it yielded the lowest confidence score among the phishing class and was classified as HIGH severity rather than CRITICAL.

Mathematical reconstruction of this specific case reveals the underlying mechanism:

- $H(e) = 1.00$ (Social engineering pattern detected)
- $B(e) = 0.20$ (Low negative sentiment probability from DistilBERT, likely due to a highly neutral or professional linguistic tone)
- $U(e) = 1.00$ (Suspicious URLs detected)

Applying the weighting coefficients ($w_1 = 0.6$, $w_2 = 0.25, w_3 = 0.15$):

$$S(e) = (0.6 \times 0.20) + (0.25\ 1.00) + (0.15 \times 1.00)$$

$$S(e) = 0.120 + 0.250 + 0.150 = 0.520$$

Given the decision thresholds T = 0.5 and T_{high} = 0.7:

Since $S(e) > T$ $(0.52 > 0.50)$, the email is correctly classified as phishing. However, because $S(e) > T_{high}(0.52 < 0.70)$, it does note reach critical severity and is a appropriately stratified as high.

This case demonstrates that the system does not blindly classify based on the mere presence of structural indicators (H and U). Instead, the dominant NLP component (B)

accurately reflects the lack of linguistic manipulation, proportionally lowering the final confidence score while still maintaining a correct binary classification. This behavior confirms the mathematical stability and forensic defensibility of the aggregation model.

4.4.4 Component Contribution Analysis

An empirical analysis of the indicator contributions across the evaluated dataset reveals that the Body-based NLP probability component B(e) is the dominant driver of the final classification score. In CRITICAL severity phishing cases, B(e) contributes an average of approximately 70–80% to the final combined score S(e), while the rule-based indicators H(e) and U(e) contribute the remaining 20–30%.

This observation aligns with the assigned weighting coefficients (w1 = 0.6) and confirms that the system relies primarily on linguistic probability for baseline classification. As noted in Section 3.9.1.a, the social engineering detector H(e) activates selectively (e.g., in only 20% of phishing samples). Therefore, H(e) and U(e) function effectively as corroborative structural signals that amplify the final score when present, rather than acting as primary standalone classifiers. Acknowledging this dominance is crucial for forensic transparency, as it clarifies that the model's high confidence in certain CRITICAL cases (such as Samples #6, #7, and #9) is derived primarily from strong linguistic manipulation signals, even in the absence of explicit URL or header anomalies.

4.5 Sample Forensic Report Output

Structured forensic reporting is a fundamental requirement in DFIR practices, as it enables investigators to document findings, justify response actions, and preserve investigative reasoning. Prior studies emphasize that effective incident response reporting should integrate analytical results with explanatory narratives (Althobaiti et al., 2021; Fairbanks, 2024)

In addition to aggregate analytical results, the DFIR process generates structured forensic reports for individual email samples. These reports consolidate analytical findings into an interpretable format suitable for incident response documentation.

Figure 4.2 shows a sample forensic analysis report generated for a phishing email. The report includes the phishing classification, assigned severity level, computed confidence score, detected indicators, and an explanatory narrative.

Email ID	1
Phishing	YES
Severity	CRITICAL
Confidence	0.9783
Social Engineering	-
URL	-
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Figure 4.2 Sample Analysis Report of a Phishing Email

Figure 4.2 reflects reporting practices observed in DFIR-oriented studies, where analytical outcomes are translated into structured documents to support accountability and post-incident review (Alaaraj & Yassin, 2024; Binnar et al., 2024).

The forensic report is generated by combining classification outcomes with indicator-level evidence. Conceptually, the severity level reported in the document is derived from the hybrid decision logic described in Section 3.9, where probabilistic linguistic output functions as the dominant classification component and rule-based indicators provide structured validation signals.

From a digital forensic perspective, the quantitative indicators generated during the analytical process (the specific B, H, and U scores) are strictly logged and traceable in the system backend. However, it is explicitly acknowledged that the current natural language explanatory narrative in the generated reports utilizes a generalized summary structure, primarily driven by the dominant classification component (typically the NLP probability). While the underlying mathematical decisions remain fully traceable and defensible, enhancing the report generation engine to produce fully dynamic, per-indicator textual explanations that perfectly mirror the specific H, B, and U values triggered for each individual email is identified as a critical improvement for future iterations to maximize human-readable interpretability.

Such structured reporting is essential in DFIR contexts because it enables justification of response actions, supports auditability, and preserves investigative reasoning.

4.6 Discussion in the Context of Incident Response

From an incident response perspective, the results demonstrate that the applied DFIR methodology effectively supports phishing incident identification and prioritisation. The clear

distinction between legitimate and phishing emails enables investigators to quickly identify high-risk incidents, while severity and confidence metrics assist in allocating response resources efficiently.

The integration of indicator-level explanations and structured forensic reporting ensures that automated analysis remains interpretable and aligned with forensic principles. This balance between automation and human judgment is critical for effective incident response operations.

From a forensic accountability perspective, the integration of severity and confidence scoring enhances structured triage during incident response. CRITICAL severity cases can be prioritized for immediate containment, while HIGH severity cases may undergo further investigation. Confidence values provide transparent justification for escalation decisions, enabling documentation of why specific response actions were taken. This interpretability is essential in audit, compliance, and post-incident review processes, where analytical reasoning must be defensible.

These results directly address the research question defined in Chapter I by demonstrating how a DFIR-based methodology can operationally support investigation and reporting of email-based phishing incidents. The structured severity differentiation facilitates triage prioritization, while the confidence score provides transparent justification for escalation decisions. The integration of probabilistic linguistic modeling with rule-based indicators ensures that analytical outcomes remain interpretable and reconstructible, thereby fulfilling the objective of structured forensic support rather than mere binary detection.

From a practical incident response perspective, the severity and confidence scoring mechanisms produced by the analytical model provide a structured basis for prioritising incident handling activities. Emails classified with higher severity levels can be immediately prioritised for containment and mitigation actions, while lower-risk emails may undergo additional verification procedures before escalation. By integrating analytical scoring with structured forensic reporting, the proposed model functions not only as a detection mechanism but also as a decision-support tool that assists incident responders in allocating investigative resources effectively.

4.7 Findings and Research Implications

Based on the results and discussion presented in this chapter, several findings can be identified. First, the DFIR-based methodology effectively distinguishes phishing emails from

legitimate communications using interpretable indicators. Second, severity and confidence metrics provide practical support for incident response prioritisation. Third, structured forensic reporting enhances transparency and accountability during investigations.

4.7.1 Model Limitations and Critical Reflection

Despite its structured and systematic design, the analytical model presents several limitations that must be critically acknowledged. First, the reliance on a sentiment-based NLP classifier introduces potential vulnerability to adversarial evasion, as sophisticated attackers may deliberately construct phishing messages using neutral or even positive linguistic tones to bypass probabilistic detection. Second, the relatively limited dataset size does not fully represent the wide diversity and evolving complexity of phishing campaigns encountered in real-world operational environments, thereby constraining generalizability. Third, the rule-based indicators depend on predefined keyword and pattern configurations, which may fail to capture newly emerging linguistic manipulation strategies or zero-day social engineering tactics. Fourth, the application of fixed probabilistic thresholds may require contextual recalibration when deployed across different organizational environments with varying risk appetites and threat landscapes. Collectively, these limitations indicate that the methodology should be positioned as a structured investigative support mechanism within a DFIR workflow rather than as a fully autonomous or definitive detection system.

The findings presented in this chapter demonstrate that the proposed DFIR-based methodology supports structured, interpretable, and operationally relevant phishing investigation. However, the analysis also reveals inherent limitations in threshold-based probabilistic modeling and rule-based indicator dependency. Future improvements should therefore focus on adaptive threshold calibration, expanded datasets, and resilience against adversarial linguistic manipulation.

4.7.2 System Performance and Operational Constraints

During the implementation and testing phases, an important operational constraint was observed regarding the forensic report generation pipeline. While the analytical aggregation and NLP inference components demonstrated high efficiency through batch processing optimizations, the final rendering of structured PDF forensic reports operates sequentially.

Empirical testing revealed that generating comprehensive, indicator-rich PDF reports for large datasets in a standard CPU-only environment incurs significant processing delays. This observation validates the methodological decision to focus the scenario-based validation on a

curated 40-sample dataset. This approach ensures that each email receives deep, traceable forensic analysis and stable report generation, rather than risking system instability or incomplete outputs through brute-force processing of massive datasets. This performance characteristic highlights that while the analytical logic is highly scalable, the reporting engine would benefit from asynchronous background processing or GPU-accelerated infrastructure in future production deployments.

CHAPTER V
CONCLUSION

5.1 Conclusion

This research investigates how a Digital Forensic Incident Response (DFIR) methodology can support the investigation and reporting of email-based phishing attacks through interpretable analytical results. The motivation for this research arises from the observation that many existing phishing detection approaches primarily focus on predictive accuracy while providing limited support for forensic interpretability, investigative traceability, and structured incident response decision-making.

The research question formulated in Chapter I asks how a DFIR-based methodology can be applied to support the investigation and reporting of email-based phishing incidents using interpretable analytical mechanisms. To address this question, this study proposes a structured analytical framework that integrates probabilistic linguistic analysis with rule-based phishing indicators derived from multiple email artifacts. The analytical model evaluates three primary indicator categories: header-based indicators associated with social engineering patterns, body-based linguistic analysis derived from probabilistic natural language processing, and URL-based structural indicators extracted from embedded hyperlinks.

The results presented in Chapter IV demonstrate that the proposed methodology provides an interpretable analytical mechanism capable of supporting phishing investigation within a DFIR context. The classification process does not rely solely on automated probabilistic outputs but instead incorporates observable indicators extracted from email artifacts. Phishing emails consistently exhibit combinations of urgency-based language, social engineering cues, and suspicious URLs. The simultaneous presence of these indicators produces higher probabilistic scores within the analytical model, causing the message to exceed the predefined classification threshold. This multi-indicator aggregation mechanism improves interpretability because classification outcomes can be directly explained through observable evidence contained within the analyzed email.

From an incident response perspective, the integration of severity and confidence scoring provides practical support for triage and prioritisation decisions. Severity levels enable investigators to identify high-risk phishing incidents that require immediate response actions, while confidence values provide transparent justification for classification outcomes

derived from probabilistic analysis. By linking probabilistic scores with structured severity differentiation, the methodology provides a consistent mechanism for translating analytical outputs into operational response decisions.

From a digital forensic perspective, the proposed approach enhances investigative transparency by ensuring that each classification decision can be traced back to identifiable email artifacts. Unlike traditional phishing detection systems that produce binary classification outputs without explanatory context, the analytical model applied in this research enables investigators to reconstruct the reasoning behind classification decisions through structured indicator analysis. This capability supports forensic accountability and strengthens the documentation of investigative findings within incident response reports.

Furthermore, this research delivers distinct scientific and practical contributions. Scientifically, it bridges the methodological gap between predictive machine learning detection and structured forensic decision-making. By formalizing a hybrid, threshold-based indicator aggregation model that combines probabilistic NLP outputs with rule-based structural signals, this study provides a reproducible and transparent framework for evaluating phishing incidents with explicit interpretability. Practically, the developed ePhish analytical support system offers incident responders a tangible decision-support tool. By translating complex analytical outputs into structured severity levels, confidence scores, and defensible forensic PDF reports, the system enhances triage efficiency, resource allocation, and accountability in real-world Security Operations Center (SOC) workflows.

5.2 Limitations

Despite the methodological contributions and analytical findings presented in this study, several limitations should be acknowledged.

First, the dataset used in this research is relatively limited, consisting of 40 email samples that include both phishing and legitimate messages. Although this dataset is sufficient for scenario-based validation within a DFIR case study context, it may not fully represent the wide diversity of phishing campaigns observed in real-world operational environments. Phishing attacks continuously evolve in terms of linguistic style, delivery mechanisms, and social engineering strategies, which may introduce variations not captured within the current dataset.

Second, the probabilistic linguistic analysis component relies on a pre-trained DistilBERT model originally designed for sentiment classification. While this model is

repurposed as a proxy heuristic indicator for linguistic manipulation, and observed results in Chapter IV show a functional separation in scores for the evaluated dataset, the direct causal link between negative sentiment polarity and malicious intent remains an assumption. This proxy approach has not been rigorously validated against adversarial, neutrally-toned, or positively-framed phishing datasets (e.g., highly sophisticated spear-phishing), which may bypass the linguistic manipulation detection mechanism.

Third, the rule-based indicator extraction mechanism depends on predefined keyword patterns associated with social engineering techniques. Although such rules provide interpretability and forensic traceability, they may fail to detect emerging phishing strategies that employ novel linguistic structures or sophisticated obfuscation techniques.

Fourth, the classification decision boundary applied in the analytical model is based on a fixed probabilistic threshold. While this threshold demonstrates stable behavior within the evaluated dataset, different organizational environments may require threshold adjustment to account for variations in communication patterns, risk tolerance, and operational context.

Fifth, the current implementation was evaluated on a standard CPU-only hardware environment. While the NLP inference component was optimized via batch processing to ensure computational efficiency, the generation of structured forensic PDF reports remains a computationally intensive and sequential process. Empirical testing revealed that generating comprehensive forensic reports for larger datasets resulted in significant processing delays due to hardware resource constraints. This limitation restricts the system's current real-time applicability for high-throughput, large-scale incident response scenarios without infrastructure upgrades.

These limitations indicate that the analytical model should be interpreted as an investigative support mechanism designed to assist incident responders and forensic analysts rather than as a fully autonomous phishing detection system.

5.3 Future Work

Several directions for future research can be proposed based on the limitations identified in this study.

First, future studies may extend the analytical framework by incorporating larger and more diverse phishing datasets collected from different organizational environments and threat intelligence sources. A larger dataset would enable more comprehensive evaluation of

the proposed methodology and provide deeper insight into the variability of phishing attack patterns.

Second, future research could develop phishing-specific natural language processing models trained directly on labeled phishing datasets rather than relying on sentiment-based proxy modeling. Such models may capture subtle linguistic manipulation strategies more accurately and improve the reliability of probabilistic classification.

Third, the integration of advanced technical indicators could further strengthen the forensic analysis process. Future iterations should implement deep header analysis, specifically parsing and validating SPF, DKIM, and DMARC records, as well as detecting routing anomalies, to strengthen the header-based component beyond basic sender extraction. Additionally, domain reputation analysis, hyperlink redirection analysis, and attachment-based threat indicators should be incorporated.

Fourth, adaptive threshold calibration techniques may be explored to dynamically adjust classification decision boundaries based on contextual factors and historical incident response data. Adaptive thresholding could improve the robustness of the analytical model when deployed across different operational environments.

Fifth, the PDF report generation engine should be upgraded to produce fully dynamic, per-indicator natural language explanations. This enhancement will ensure that the textual narrative in the forensic report perfectly mirrors the specific H, B, and U values triggered for each individual email, thereby eliminating generalized summaries and maximizing human-readable interpretability for incident responders.

Finally, to address the computational bottlenecks observed during forensic report generation and facilitate operational deployment, future iterations should implement asynchronous background processing or leverage GPU-accelerated infrastructure. Decoupling the analytical inference pipeline from the report rendering engine will significantly reduce processing latency, enabling the system to handle large-scale datasets efficiently. Building upon these technical enhancements, future research may investigate the integration of the proposed DFIR-based analytical framework into operational incident response platforms or Security Operations Center (SOC) environments. Evaluating the methodology within real-world incident response workflows would provide valuable insights into its effectiveness in supporting phishing investigation, triage prioritisation, and forensic reporting in practical cybersecurity operations.

BIBLIOGRAPHY

Alaaraj, A., & Yassin, A. (2024). *Degree project Investigating Cybersecurity Response Strategies: Measures to Responding to Successful Spear Phishing Attacks Bachelor Degree Project in Informatics First Cycle 30 credits Spring term 2024*. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1878451

Alsubaei, F. S., Almazroi, A. A., & Ayub, N. (2024). Enhancing Phishing Detection : A Novel Hybrid Deep Learning Framework for Cybercrime Forensics. *IEEE Access*, *12*(December 2023), 8373–8389. https://doi.org/10.1109/ACCESS.2024.3351946

Althobaiti, K., Jenkins, A. D. G., & Vaniea, K. (2021). A Case Study of Phishing Incident Response in an Educational Organization. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2). https://doi.org/10.1145/3476079

Altulaihan, E., Alismail, A., Rahman, M. M. H., & Ibrahim, A. A. (2023). *Email Security Issues , Tools , and Techniques Used in Investigation*. 1–28. https://doi.org/10.3390/su151310612

Asliyuksek, H., & Tonkal, O. (2025). *A Comparative Evaluation of a Multimodal Approach for Spam Email Classification Using DistilBERT and Structural Features*. 1–36. https://doi.org/10.3390/electronics14193855

Banu, S., Divya, R., Tt, D. D., Sri, B., Gheisari, M., Khammar, S., & Ghaderzadeh, M. (2025). Phishing Attack Simulation, Email Header Analysis, and URL Scrutiny: A Comprehensive Approach to Cyber Threat Mitigation. *Computer Networks and Communications*, *3*(2), 1–20. https://doi.org/10.37256/cnc.3220256797

Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., & Gollangi, H. K. (2024). *AI-Driven Phishing Email Detection : Leveraging Big Data Analytics for Enhanced Cybersecurity*. *44*(3), 7211–7224. https://doi.org/10.2139/ssrn.5029438

Binnar, P., Bhirud, S., & Kazi, F. (2024). Security analysis of cyber physical system using digital forensic incident response. *Cyber Security and Applications*, *2*(July 2023). https://doi.org/10.1016/j.csa.2023.100034

Burda, P., Allodi, L., Serebrenik, A., & Zannone, N. (2024). "Protect and Fight Back": A Case Study on User Motivations to Report Phishing Emails. *ACM International Conference Proceeding Series*, 30–43. https://doi.org/10.1145/3688459.3688473

Burita, L., Matoulek, P., Halouzka, K., & Kozak, P. (2021). Analysis of phishing emails. *AIMS Electronics and Electrical Engineering*, *5*(1), 93–116. https://doi.org/10.3934/ELECTRENG.2021006

Butt, U. A., Amin, R., Aldabbas, H., Mohan, S., Alouffi, B., & Ahmadian, A. (2023). Cloud-based email phishing attack using machine and deep learning algorithm. *Complex & Intelligent Systems*, *9*(3), 3043–3070. https://doi.org/10.1007/s40747-022-00760-3

Damatie, E. M., Eleyan, A., & Bejaoui, T. (2024). Real-Time Email Phishing Detection Using a Custom DistilBERT Model. *2024 International Symposium on Networks, Computers and Communications (ISNCC)*, 1–6. https://doi.org/10.1109/ISNCC62547.2024.10759011

Dimitriadis, A., Lontzetidis, E., Kulvatunyou, B., Ivezic, N., Gritzalis, D., & Mavridis, I. (2023). Fronesis : Digital Forensics-Based Early Detection of Ongoing Cyber-Attacks. *IEEE Access*, *11*(January), 728–743. https://doi.org/10.1109/ACCESS.2022.3233404

Fairbanks, J. (2024). *POST: Email Archival, Processing and Flagging Stack for Incident Responders*. https://doi.org/10.48550/arXiv.2407.01433

Kuforiji, J. (2025). *Digital Forensics and Incident Response (DFIR) Automation : Leveraging AI to Accelerate Breach Investigation , Evidence*. *01*(04). https://doi.org/10.64235/tsvfvz27

Muneer, A., Ali, R. F., Al-Sharai, A. A., & Fati, S. M. (2021). A Survey on Phishing Emails Detection Techniques. *4th International Conference on Innovative Computing, ICIC 2021*, *Icic*, 0–5. https://doi.org/10.1109/ICIC53490.2021.9692960

Nailah, S., & Rosnelly, R. (2025). *Analisis Ancaman Phishing Di Platform Media Sosial Email menggunakan Metode Digital Forensic Research Workshop*. *3*(2), 450–461. https://kti.potensi-utama.org/index.php/JUREKSI/article/view/2168

Purnamasari, & Sutabri, T. (2023). *Analisis kejahatan online phising pada institusi pemerintah / pendidik sehari-hari*. *06*(01), 29–34. https://doi.org/10.32502/digital.v6i1.5620

Putra, F. P. E., Ubaidi, Zulfikri, A., Arifin, G., & Ilhamsyah, R. M. (2024). *Analysis of Phishing Attack Trends , Impacts and Prevention Methods : Literature Study*. *4*(1), 413–421. https://doi.org/10.47709/brilliance.v4i1.4357

Ramadhani, E., & Raharjo, T. (2025). *Enhancing Digital Forensics with Cyber Kill Chain and 5W1H : A Case Study on Phishing Attacks*. *11*(1), 79–89. https://doi.org/doi.org/10.21108/ijoict.v11i1. 8966

Rastenis, J., Ramanauskaitė, S., Suzdalev, I., Tunaitytė, K., Janulevičius, J., & Čenys, A. (2021). Multi-language spam/phishing classification by email body text: Toward automated security incident investigation. *Electronics (Switzerland)*, *10*(6), 1–10.

https://doi.org/10.3390/electronics10060668

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2022). A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques. *IEEE Access*, *10*, 65703–65727. https://doi.org/10.1109/ACCESS.2022.3183083

Selamat, S. R., Rizal, R., Nursihab, C., & Amien, N. (2024). Advanced Phishing Attack Detection Through Network Forensic Methods and Incident Response Planning Based on Machine Learning. *International Journal of Informatics and Computing*, *1*(1), 19. https://iaico.org/index.php/JICO/article/view/1/3

Simoiu, C., Zand, A., Thomas, K., & Bursztein, E. (2020). *Who is targeted by email-based phishing and malware.pdf*. https://doi.org/doi.org/10.1145/3419394.3423617

Stojnic, T., Vatsalan, D., & Arachchilage, N. A. G. (2021). *Phishing email strategies : Understanding cybercriminals ' strategies of crafting phishing emails*. *May 2020*, 1–17. https://doi.org/10.1002/spy2.165

Zakiyaturrahma, & Riadi, I. (2022). *Email Forensic from Phishing Attack using Network Forensics Development Life Cycle Method*. *183*(46), 36–42. https://doi.org/10.5120/ijca2022921865

APPENDIX

Appendix Data Phishing Email

ePhish – Digital Forensic Email Analysis Report

Email Evidence #1

Email ID	1
Phishing	YES
Severity	CRITICAL
Confidence	0.9842
Social Engineering	-
URL	http://whitedone.com/
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #2

Email ID	2
Phishing	YES
Severity	CRITICAL
Confidence	0.9642
Social Engineering	-
URL	http://www.brightmade.com
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #3

Email ID	3
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #4

Email ID	4
Phishing	YES
Severity	CRITICAL
Confidence	0.9998
Social Engineering	confirm
URL	http://en.wikipedia.org/wiki/.so_%28domain_name%29 http://svn.apache.org/viewvc?rev=619753&view=rev http://svn.apach
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #5

Email ID	5
Phishing	YES
Severity	CRITICAL
Confidence	0.9666
Social Engineering	-
URL	http://7iwfna.blu.livefilestore.com/y1pXdX3kwzhBa8xhXv8tdHbjHn7Tj4VT91YQg5lB5-_H
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #6

Email ID	6
Phishing	YES
Severity	CRITICAL
Confidence	0.9949
Social Engineering	-
URL	-
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #7

Email ID	7
Phishing	YES
Severity	CRITICAL
Confidence	0.998
Social Engineering	-
URL	-
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #8

Email ID	8
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #9

Email ID	9
Phishing	YES

Severity	CRITICAL
Confidence	0.9991
Social Engineering	-
URL	http://issues.apache.org/SpamAssassin/show_bug.cgi?id=5780
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #10

Email ID	10
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #11

Email ID	11
Phishing	YES
Severity	CRITICAL
Confidence	0.9955
Social Engineering	bank
URL	http://freshplateworks.com
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #12

Email ID	12
Phishing	YES
Severity	CRITICAL
Confidence	0.9893
Social Engineering	-
URL	http://oemmerchant.com
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #13

Email ID	13
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #14

Email ID	14
Phishing	YES

Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #15

Email ID	15
Phishing	YES
Severity	CRITICAL
Confidence	0.997
Social Engineering	-
URL	http://pacelike.com
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #16

Email ID	16
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://ieinternet.com/mailwall/) http://ieinternet.com/mailwall/)
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #17

Email ID	17
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #18

Email ID	18
Phishing	YES
Severity	CRITICAL
Confidence	0.9961
Social Engineering	-
URL	http://www.cnn.com/video/partners/email/index.html?url=/video/us/2008/07/31/moos http://www.cnn.com/video/partners/email
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #19

Email ID	19

Phishing	YES
Severity	CRITICAL
Confidence	0.9993
Social Engineering	suspend, security
URL	http://lists.tuxonice.net/mailman/listinfo/tuxonice-users http://lists.tuxonice.net/mailman/listinfo/tuxonice-users http
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Email Evidence #20

Email ID	20
Phishing	YES
Severity	HIGH
Confidence	0.6691
Social Engineering	immediately
URL	https://list.scms.waikato.ac.nz/mailman/listinfo/wekalist https://list.scms.waikato.ac.nz/pipermail/wekalist/attachments
Explanation	Email diklasifikasikan sebagai phishing karena mengandung pola social engineering, urgensi tinggi, dan konteks manipulatif.

Appendix Data Legitimate Email

ePhish – Digital Forensic Email Analysis Report

Email Evidence #1

Email ID	1
Phishing	NO
Severity	LOW
Confidence	0.993
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #2

Email ID	2
Phishing	NO
Severity	LOW
Confidence	0.9932
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #3

Email ID	3
Phishing	NO
Severity	LOW
Confidence	0.995
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #4

Email ID	4
Phishing	NO
Severity	LOW
Confidence	0.9953
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #5

Email ID	5
Phishing	NO
Severity	LOW
Confidence	0.9945
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #6

Email ID	6
Phishing	NO
Severity	LOW
Confidence	0.996
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #7

Email ID	7
Phishing	NO
Severity	LOW
Confidence	0.9941
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #8

Email ID	8
Phishing	NO
Severity	LOW
Confidence	0.9939
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #9

Email ID	9
Phishing	NO
Severity	LOW

Confidence	0.9955
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #10

Email ID	10
Phishing	NO
Severity	LOW
Confidence	0.9906
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #11

Email ID	11
Phishing	NO
Severity	LOW
Confidence	0.9916
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #12

Email ID	12
Phishing	NO
Severity	LOW
Confidence	0.9932
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #13

Email ID	13
Phishing	NO
Severity	LOW
Confidence	0.9927
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #14

Email ID	14
Phishing	NO
Severity	LOW

Confidence	0.9939
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #15

Email ID	15
Phishing	NO
Severity	LOW
Confidence	0.9921
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #16

Email ID	16
Phishing	NO
Severity	LOW
Confidence	0.9931
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #17

Email ID	17
Phishing	NO
Severity	LOW
Confidence	0.9915
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #18

Email ID	18
Phishing	NO
Severity	LOW
Confidence	0.9933
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #19

Email ID	19
Phishing	NO
Severity	LOW

Confidence	0.994
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.

Email Evidence #20

Email ID	20
Phishing	NO
Severity	LOW
Confidence	0.9949
Social Engineering	-
URL	-
Explanation	Email tidak menunjukkan indikator phishing signifikan.