

# Smart Detection and Prevention of Whaling Attacks in Cyber Security

Dr. Ann Zeki Ablahd Magdacy Jerjes<sup>1\*</sup>, Ihsan Hassan Hussein<sup>2</sup>, and  
Dr. Israa Tahseen Ali-Attar<sup>3</sup>

<sup>1</sup>Department of IT and Computer Network, Technical Engineering College for Computer and AI,  
Northern Technical University, Mosul, Iraq. drann@ntu.edu.iq,  
<https://orcid.org/0000-0002-8501-6540>

<sup>2</sup>Department of IT and Computer Network, Technical Engineering College for Computer and AI,  
Northern Technical University, Mosul, Iraq. ihsan.bayoglu@ntu.edu.iq,  
<https://orcid.org/0009-0002-5607-4474>

<sup>3</sup>Department of IT and Computer Network, Technical Engineering College for Computer and AI,  
Northern Technical University, Mosul, Iraq. israa.24ali@ntu.edu.iq,  
<https://orcid.org/0000-0001-5191-9349>

Received: September 25, 2025; Revised: October 31, 2025; Accepted: December 24, 2025; Published: February 27, 2026

## Abstract

Whaling attacks is a sophisticated spear phishing form of targeting high-level executives. It poses a significant threat to the organization's cybersecurity. This research presents and introduces a novel intelligent system, innovative in detecting and preventing all the complex threats of enterprise environments. The power of artificial intelligence with Python implements a system that uses Natural Language Processing (NLP) to analyze the contents of emails for identifying the subtle indicators of malicious intent through suspicious patterns, sentiment analysis, and contextual cues. The performance of the proposed system was high because a hybrid model of machine learning by using a Support Vector Machine (SVM) classifier, with integrating TF-IDF vectorization, was used to enhance the accuracy of detecting malicious email data. The comprehensive evaluation in a real-world dataset for whaling attacks demonstrates the efficacy of an AI-driven approach by achieving exceptional results with a precision of 99.2%, a recall of 98%, and an F1-score of 98.596%. Furthermore, the proposed intelligence system is novel because it incorporates real-time threat detection to prevent effective capabilities in whaling attacks by minimizing false positives and accelerating response times. This system will deliver an essential contribution to the cybersecurity practices, offering an adaptive solution that fortifies enterprise defenses against increasingly sophisticated social engineering attacks, and for reducing the financial fraud risk and data breaches.

Keywords: SVM, NLP, Whaling Attacks, Email Security, Spear Phishing, Threats, ML, Cyber Security.

## 1 Introduction

The increasing challenge of the contemporary enterprise landscape is sophisticated and targeted cyber threats, among which whaling attacks have emerged as a particularly insidious form. In contrast, the traditional phishing campaigns (Mokoena & Nilsson, 2023; Oliveira et al., 2021) employed a broad network approach, like whaling, targeting high-level executives, including CEOs, CFOs, and other decision makers, by primary exfiltration of sensitive objectives for orchestrating fraudulent financial transactions and all organizational data (Olaniyan, 2024). The intrinsic value and power associated with these targets make whaling attacks highly destructive and extremely difficult to pinpoint. The attackers continuously improve their social engineering techniques while traditional security systems demonstrate their ineffectiveness in spam filter and static rule-based detection systems (Vembarasi et al., 2024; Kaushik, 2023; Kalech, 2019). The current prevailing legacy detection systems are unable to distinguish between genuine and fraudulent correspondence because attackers create personalized content that looks authentic to business messages (Nguyen et al., 2023). This research develops an innovative framework that targets whaling attacks within enterprise environments (Huang et al., 2023; Chang et al., 2013) to fix their significant limitations. The proposed system used a combination of natural language processing (NLP) and the algorithm of machine learning (ML) in detecting malicious email content. The strengthened enterprise of the cybersecurity proposed system appears through the threat in monitoring real-time, while it will actively prevent security threats. The study demonstrates the necessity of implementing adaptive security measures with intelligence features to fight whaling attacks by providing flexible defense solutions for contemporary business operations. This research is contributed:

- In developing a novel framework with AI-driven methods for preventing and detecting whaling attacks by using a hybrid integration of NLP techniques and an SVM classifier.
- In using a mathematical approach of modeling that formalizes the problem of detecting email-based whaling attacks.
- In addition to that, it implements a module capable of identifying suspicious behavior of all threats in real-time.
- It contributes to achieving superior performance metrics by including a recall 98%,98.596% of F1 score, and 99.2% precision of outperforming traditional rule-based and spam-filtering mechanisms.

The given paper presents an AI-based detector and prevention system of whaling attacks on high-level executives. I: Introduction is the background section that sheds light on the issue, including the shortcomings of the classical security systems and the AI-based solution to the problem. Section II: Related Works presents the current research in the area of phishing attacks and whaling attacks that indicates the necessity of more sophisticated detection systems. Section III: Structure of the Proposed System provides the modular design of the system that incorporates input: Email, preprocessing, feature extraction by NLP, and SVM classification. Part IV: Features of Whaling discusses the peculiarities of the whaling attacks. Section V: Email Structure provides a discussion on the fact that email headers, bodies, and attachments can be analyzed to identify malicious intent. Section VI: Natural Language Processing (NLP) reveals the application of NLP in analyzing the email content. Section VII: TF-IDF Vectorization provides information about the usage of this technique to extract features, enhancing the accuracy of the classification. Section VIII: Support Vector Machine Algorithm explains the SVM algorithm that was used in email classification. Section IX: Analysis of Results gives the performance metrics of the system, which are precision, recall, and F1-score. Section X: Discussion: The discussion

of the results and the effectiveness of the system. Lastly, Section XI: Conclusion presents a conclusion of the main findings and recommendations for further research.

## 2 Related Works

The increased level of cybersecurity threats has seen more investigations into the issue of whaling attacks, and this has served to provide the much-needed multi-layered protective systems that are highly advanced. Such laser spear-phishing gears that target chief officers have quite severe implications because it is highly individualized, making it extremely dangerous in terms of their ability to cause substantial financial and reputational losses. Initial studies of (Olaniyan, 2024) indicated how the executive was a target of social engineering attacks, which led to executive-specific training, AI simulations, and behavior authentication training. Similarly, (Vanitha et al., 2024) the whaling attacks are effective, and their success mainly depends on their personalized phishing techniques, calling to introduce context-aware systems, special AI algorithms, and external threat analysis. Effective whaling operations have a devastating impact on organizations, as pointed out by (Harris & Molzahn, 2024), necessitating tailored cyber-insurance solutions and rapid response protocols

The classical solutions, including fixed rule-based systems and conventional spam filters, have shown poor results in blocking such advanced malware and threats, as revealed (Rajan & Srinivasan, 2025). They insist on strong security levels, which incorporate AI-based email protection, Multi-Factor Authentication (MFA), which can be adjusted, and joint threat intelligence sharing. In the same way, (Birthriya et al., 2025) accentuated the supportive role of the executive-specific training and the monitoring of the ongoing communication. Recent developments, however, deal with the increasing sophistication of these attacks. Emphasized the importance of expedited creation of deep fake identifiers and multi-national defense systems, as the tactics of high-tech social engineering increased in scope (Guru Prasad & Badrinarayanan, 2025). An approach as a multi-solution with particular whaling defense procedures and a high-tech AI analytics system (Saleh et al., 2024). Moreover, modern cybersecurity research activities also show the validity of the use of Quantum-proof encryption, human behavior monitoring systems, zero-trust infrastructure, and AI analytics to secure such complex technology as blockchain (Yeoh et al., 2022). The growing complexity of cybersecurity has made the study of whaling attacks even more complicated, which is why a sophisticated and multi-layered protection is necessary. These highly specific spear-phishing operations, targeting senior executives, pose a significant risk because of their personal character and the fact that they may cause substantial financial and reputational losses.

It is observed that in the literature review, whaling attacks, which are directed against top executives, have become more complex and costly to both companies in terms of finances and reputation. Conventional security systems, such as spam filters, do not work in such customized attacks. It has been proven that whaling attacks are successful, which has resulted in the creation of sophisticated defenses, such as AI-based, multi-factor authentication (MFA), executive-specific training, and deep fake detection. In order to achieve mitigation of these risks effectively, multi-layered and dynamic security mechanisms with AI, human behavior tracking, and the latest encryption tools must be used by organizations.

## 3 Structure of the Proposed System

Figure 1 illustrates the workflow of the machine learning-based system for whaling attack detection and predictive maintenance. This is initiated by Data Sources (email data or sensor data), which are then

followed by Preprocessing steps, which purify and normalize the data. The next step is the Feature Engineering that mines out the relevant features and feeds them to the Machine Learning Models to train them. The Model Training and Evaluation stage is an evaluation of the models. After training, the system moves to Real-Time Detection, where predictions are made on incoming data. Finally, the Output provides alerts or actions based on the classification results (e.g., quarantining emails or scheduling maintenance). The arrows indicate the flow of data and actions across these stages.

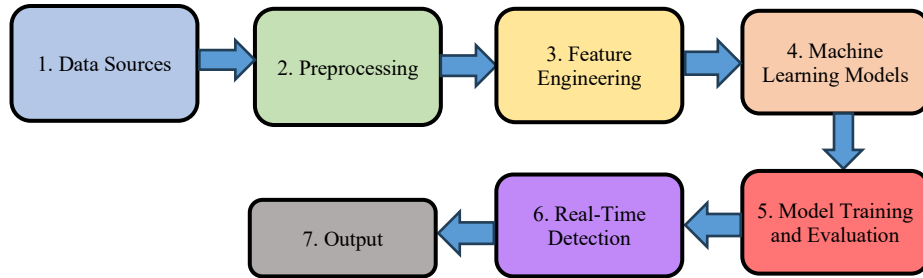


Figure 1: Architecture diagram for whaling attack detection and predictive maintenance system

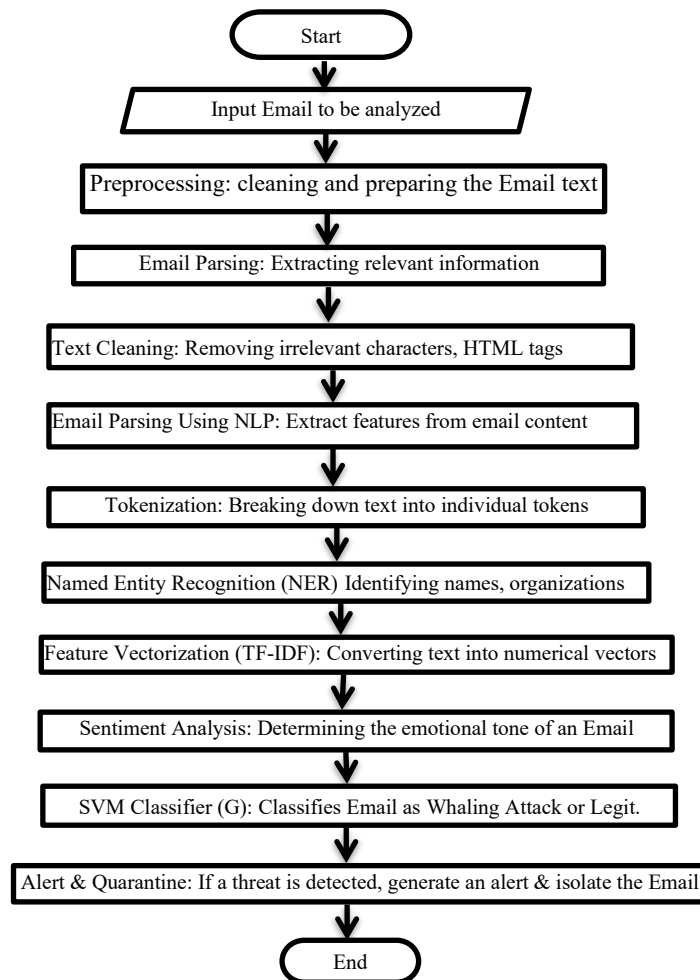


Figure 2: Structure of the proposed system

The proposed system employs a modular architecture for intelligent detection and prevention of whaling attacks (Mironeanu et al., 2021). A modular architectural design enables this system to detect

and stop whaling attacks effectively (Mironeanu et al., 2021). The system has two initial components, which consist of an Email Input Module and a Preprocessing Module. A real-time threat intelligence database tests the emails while an SVM model (Kandhro et al., 2025; Senthilkumar et al., 2023) obtains classification results (Ablahd et al., 2024). The system requires a threat match to activate an alert and email quarantine, but no threats lead to delivery with logging of classification results. The proposed system structure contains its components as shown in Figure 2.

### **Phishing and Whaling Distinctions Techniques**

Phishing represents a cyber-attack approach that deceives people into sharing their private data. The modern form of technical crime occurs when attackers exploit social engineering for deception. The primary delivery method for phishing attacks occurs through several deceptive communication networks, which include both Email and text messaging (smishing), in addition to phone calls (vishing) and social media platforms. Attack methods employed in phishing involve spoofing, as well as subjecting victims to scare campaigns and the practice of pretexting combined with harmful file attachments.

### **Common Types of Phishing**

Phishing is a variety of dishonest methods fraudsters use to fool customers into disclosing non-public information or downloading dangerous programs. Every type of phishing attack has precise trends and objectives that let attackers personalize their procedures to target specific victims or achieve particular goals.

### **Spear Phishing**

This is an entirely focused kind of phishing in which the assailant creates custom-designed communications intended for a particular character or agency. Unlike mass-sent popular phishing emails, spear phishing is primarily based on thorough research of the goal to boost credibility, including the usage of the recipient's call, painting's role, or latest pastime. Usually posing as trusted contacts, those attacks use manipulation of the sufferer into clicking risky hyperlinks or revealing non-public statistics.

### **Whaling**

A specialist form of spear phishing, whaling is based on high-profile targets, consisting of CEOs, CFOs, or different pinnacle executives. Usually disguising themselves as crucial enterprise communications, including criminal subpoenas, economic reviews, or messages from regulatory government, these assaults are supposed to take advantage of top-level employees' energy and access level.

### **Clone Phishing**

Attackers in clone phishing copy a formerly accurate and trustworthy email, usually one the victim has already acquired, then resend it with changed textual content or harmful links. Using the familiarity and meant protection of the original message, the purpose is to fool the receiver into beginning an inflated attachment or hyperlink that seems precisely just like the original. Spear and whaling, along with clone phishing, belong to the standard classification of phishing attacks, followed by credential harvesting phishing. The attackers use fake login portals as part of credential harvesting phishing to steal user login information. The Whaling phishing stands as a specific type of spear phishing targeting senior executive personnel.

## 4 Characteristics of Whaling

The Whaling attack is a specific form of phishing that targets upper-level executives and top-level personnel at organizations. Specific individuals become the targets in whaling attacks since attackers use personalized information to boost their chances of deception. The attackers use social engineering methods to force unsuspecting targets into sharing confidential data while authorizing monetary funds. Various approaches exist for carrying out Whaling Attacks, with the following techniques among them, such as Email Spoofing, Malicious Links and Attachments, and Social Engineering. In Email Spoofing, the attackers will send emails that appear to come from legitimate sources within the organization, but in fact, they come from an unknown source, often mimicking the style and tone of senior executives. In social Engineering, for leveraging personal information about the target, attackers will create a sense of importance or urgency that forces the victim to act quickly without due diligence. But the malicious links and attachments mean that the whaling attacks do not rely on obvious malicious links or attachments; some may include links to fake websites designed to harvest credentials or install malware. There are various threats posed by whaling attacks, such as Financial Loss, which can result in significant monetary losses due to unauthorized wire transfers or fraudulent transactions. The FBI reported two years ago that over \$1.2 billion in business losses were attributed to such attacks. While a Data breach is another threat, Sensitive corporate information, including trade secrets and client data, can be compromised, leading to long-term reputational damage. The last threat is Operational Disruption. This type will disrupt all operations of the business, causing loss of productivity, downtime, and as organizations scramble to mitigate the damage.

## 5 Email Structure

The detection of a Whaling attack necessitates a comprehensive email analysis by verifying the authoritative addresses of the sender and also tracing Email through the received header and identifying unusual routing, spoofing domains, and testing metadata.

### **Email Headers Analysis**

The proposed system of Whaling attack detection critically analyzes the headers of emails by scrutinizing the sender's address for verifying the discrepancies between actual sending servers and displayed, spoofing, and examining domain variations. Also, it traces the paths of email for a received header to a malicious server and suspicious routing, testing Message-ID, communication domain, and anomaly dates indicative of messages—all these steps aimed to attempt to impersonate trusted entities.

### **Email Body Analysis**

The proposed system of whaling attack detection analyzed the Email's content meticulously by encompassing language. There are different types of analysis. The first is contextual analysis to determine request legitimacy, the second is sentiment analysis to detect emotional cues like urgency or authority. Further scrutiny involves Named Entity Recognition to classify entities and verify their consistency with known contacts. The Python library (TF-IDF), which is a machine learning pattern recognition, is used for detecting writing style, deviations from the sender's established style, authenticity, and comprehensive assessment of the Email's intent.

### **Attachment Analysis**

The proposed system of whaling attack detection analyzes the attachment file for identifying suspicious disguises and scans it to detect malware using sandboxing techniques and malware signatures. The analysis is done within a controlled environment to mitigate the risk of malicious payload execution.

### Link and Behavioral Analysis

The whaling attack detection system analyzed the links. The analysis of links involves misspelled domains, all embedded URLs, malicious sites or patterns, and analyzing the linked domain's reputation. The destination of each link is compared and verified to ensure the legitimacy and safety of the content link. The behavioral analysis is done with this proposed system by testing communication patterns for flagging unusual requests or changes in communication style, conducting time-based analysis to detect suspicious timing patterns, and identifying significant deviations from a target's typical behavior. The SVM algorithm is used to classify emails as "whaling attack" or "benign" based on the extracted features (Hajiali et al., 2019). This system must be accurate and adaptive.

### Mathematical Model for Whaling Attack Detection

Let  $X = \{x_1, x_2, \dots, x_n\}$  be the set of input email samples, where each email  $x_i$  is transformed into a TF-IDF feature vector  $v_i \in \mathbb{R}^m$ , representing the textual features of the Email.

#### Labels

Let  $y_i \in \{0,1\}$  denote the label of each Email, where:

- $y_i = 0$  indicates a benign email.
- $y_i = 1$  indicates a whaling email.

#### SVM Decision Function

The classifier learns a decision function of the form:

$$f(v_i) = \text{sign}(w \cdot v_i + b) \quad (1)$$

In Equation 1, where:

- $w$  is the weight vector, which is learned during the training process.
- $v_i$  is the feature vector of the  $i$ -th Email.
- $b$  is the bias term.

#### SVM Optimization Problem

The SVM optimization problem is formulated as:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad (2)$$

In Equation 2, where:

- $C$  is a regularization parameter that controls the trade-off between maximizing the margin and minimizing classification errors.
- $\xi_i$  are slack variables that allow for some misclassification, ensuring that the model can handle non-linearly separable data.
- The constraints are:

$$y_i(w \cdot v_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

These constraints ensure that the margin between the two classes is maximized while minimizing the errors.

### TF-IDF Mathematical Description

**Term Frequency (TF):** Measures how frequently a term appears in a document relative to other terms.

$$TF(t, d) = \frac{f(t, d)}{\sum_k f(k, d)} \quad (3)$$

In Equation 3, where:

- $f(t, d)$  is the frequency of term  $t$  in document  $d$ .
- The denominator is the total number of terms in document  $d$ .

**Inverse Document Frequency (IDF):** Measures the importance of a term within the entire corpus of documents

$$IDF(t) = \log \left( \frac{N}{n_t} \right) \quad (4)$$

In Equation 4, where:

- $N$  is the total number of documents.
- $n_t$  is the number of documents that contain the term  $t$ .

**TF-IDF:** The product of Term Frequency (TF) and Inverse Document Frequency (IDF) gives the TF-IDF score for each term (equation 5):

$$TF-IDF(t, d) = TF(t, d) \times IDF(t) \quad (5)$$

These metrics highlight essential terms in the Email by considering their frequency within a document and their rarity across the entire dataset.

## 6 Natural Language Processing (NLP)

The (NLP) Natural Language Processing, as a branch of artificial intelligence, focuses on the creation of ways in which the computer and the human language can interact. Machines, through this technology, can process human language with both comprehension and application of meaning. NLP contains various workloads that encompass text processing for tokenizing text elements, sentence breakdown, and the creation of natural language text generation and analysis for understanding written content. Through NLP technology, organizations accomplish linguistic pattern detection along with email header examination while identifying suspicious phrases, emotional content, and the underlying motives in text. The NLP algorithm handles data extraction and cleansing, followed by text translation, and stops word elimination and extraction of features from the text.

### NLP algorithms

The steps of using NLP algorithms for analyzing the contents of emails to know the suspicious language, sentiment, and intent to identify phishing attempts, and particularly whaling attacks are:

1. **Data Collection and Preparation:** At first, the data of emails was gathered for both phishing and legitimate emails (including whaling examples) to train and test the built NLP proposed system.
2. **Cleaning Data:** Removing all HTML tags and irrelevant formatting in email contents.
3. **Text Decoding:** Ensure the text for appropriately decoded and for using Unicode or entities for HTML pages.

4. Remove all stopping words: NLTK library in Python (Abro et al., 2025) code were used to eliminate all common stopping words like ('a', 'is', 'the',). And replace or remove symbols and special characters.
5. Extract all Features: Through this step text tokenization done by breaking down the email text into tokens (individual words). Consider n-grams (An n-gram is a contiguous sequence of n words or characters from a given text. In email analysis, n-grams are commonly used in the proposed system for text feature extraction, which helps in detecting phishing or whaling attempts. For sentiment analysis it used Scikit-learn tools in Python (Sukavatee & Sutthiroj, 2025). TensorFlow/PyTorch is a Python deep learning frameworks tool used in training the proposed system.

## 7 TF-IDF Vectorization

The Term Frequency-Inverse Document Frequency (TF-IDF) is a crucial technique in analyzing text using machine learning. The Whaling attack is a phishing attack that is sophisticated targeting high-profile individuals. It requires an advanced methods of detection for preventing reputational damage and significant financial. There are different terms used in working with TF-IDF Vectorization such as Term Frequency (TF), Inverse Document Frequency (IDF), TF-IDF Calculation. The TF used for calculating the number of times that appears a term in a document divided by total number of document terms. IDF, it calculates the total number of documents by the number of documents containing term. The score TF-IDF is a product of IDF and TF providing a measure of weighting of the critical term in a document relative with entire corpus.

The tool provided by the Scikit-learn (sklearn) library in Python that used in feature extraction technique in the field of Natural Language Processing (NLP) and machine learning TfidfVectorizer. It is used in identifying the words that are most indicative of malicious emails (Jagatic et al., 2007). The words that are common in phishing emails but rare in regular emails will have high TF-IDF scores. That allows the proposed system in focusing for the most relevant features and improves its accuracy. (Jyothi et al., 2024) There are different parameters and features in Python used in detection whaling attacks like:

- `max_df`: Ignores terms that have a document frequency strictly higher than the given threshold (corpus-specific stop words).
- `min_df`: Ignores terms that have a document frequency strictly lower than the given threshold.
- `max_features`: used for building a vocabulary that only considers the top `max_features` ordered by term frequency across the corpus.
- `ngram_range`: Specifies the lower and upper boundary of the range of n-values for different word n-grams or sequences of words to be extracted.
- `stop_words`: Used for removing the common words like "a," "the," and "is" that is not important.
- `vocabulary`: Accepts a predefined vocabulary.

### TF-IDF in Whaling Attack Detection

There are different steps done in using TF-IDF Vectorization like features extractions used to extract features from web content or emails. It helps for identifying phrases or keywords that are more indicative of whaling attacks, such as company names, urgent language and titles. Other steps are machine learning integration algorithm (SVM) is used for classifying legitimate or malicious emails. The last one is real time detection; the proposed system can quickly analyze incoming emails and flag potential whaling attacks before they reach their targets.

There are different reasons for using TF-IDF in Whaling Attack Detection such as improving the accuracy by focusing on the most relevant terms in the context of whaling attacks, for reducing the dimensionality of the feature space, making the model more efficient and less prone to over fitting and TF-IDF can be easily adapted to new types of whaling attacks by updating the training dataset with new examples because the sophisticated attackers may use evasion techniques such as images or audio to bypass text-based detection methods. This system requires to be updated continuously by a new data to be effective against whaling attack.

## 8 Support Vector Machine Algorithm

The Support Vector Machines (SVM) shows itself as a reliable supervised machine learning algorithm which continuously delivers high effectiveness for classification purposes. Detection of whaling attacks depends on categorizing malicious and benign emails so SVM demonstrates essential value in classification tasks. The SVM demonstrates two beneficial capabilities which are data dimension management alongside robust performance thus making it an excellent candidate for this application. The sophisticated Whaling attack requires superior detection techniques due to its social engineering characteristics. The system implements SVM to evaluate email characteristics which result in classification for determining malicious or benign behavior to boost detection rates. The choice of SVM exists due to its proven strength with high-dimensional data.

### Data Preprocessing in SVM Algorithm

The dataset of this proposed system consists of emails, their header data and textual content that obtained from Kaggle (<https://www.kaggle.com/datasets>). The original Enron Email Dataset offers an extensive collection of 651,191 URLs with three subgroups of 428103 benign or secure URLs and separate sections for 96457 defacement and 94111 phishing and 32520 malware URLs. The system employed a selection of Kaggle data which incorporated simulated whaling attack email sets (Liu et al., 2024; Ramachandran et al., 2024). Voluntary contributions on the Kaggle platform considerably enhance whaling attack studies because it offers basic phishing and harmful URL datasets for building strong detection methods and it supports algorithm enhancement through public competitions and sharing notebooks and it emphasizes in essential feature engineering practices including sentiment and linguistic analysis. The collaborative atmosphere helps multiple users exchange knowledge while receiving community feedback which promotes general phishing defense and anomaly detection methods and operates as an essential platform for improving cyber security research and innovation.

Rarity of direct whaling attack datasets on Kaggle allows researchers to build cross-purpose phishing detection models which adapt to whaling attack detection and enhance anomaly detection methods for executive communication analysis while the platform contributes to overall cyber security development. Kaggle is used in this cyber security system because it gives researchers the fundamental resources which support their development of robust detection and prevention systems against phishing and whaling attacks. A complete dataset was produced through simulated whaling attack emails that followed genuine social engineering methods combined with practical examples. The Python script generated simulated emails through an auto-generated process which randomly changed attributes including sender details and email text and recipient profiles. The simulated program duplicated authentic whaling attacks by using realistic elements such as demanding text and requests for crucial information in its design the proposed tool developed for producing a phishing dataset and whaling emails.

The email alteration tool enables users to revise standard email properties including sender details along with subject lines and main content. The generated email messages followed the modern standards that exist in social engineering assaults. The analysts conducted data analysis on an anonymized dataset because privacy restrictions demanded the removal of all sensitive information. SVM algorithm required a numerical transformation of all Email data inputs. The key elements for analysis include text-based content examination for email bodies as well as header domain investigation alongside sentiment analysis of tone and metadata evaluation of receiver totals and message timestamps and duration. The preprocessing of data included handling missing values combined with normalizing features while removing all unnecessary characters. The proposed system evaluation needs the data to be divided into testing and training subsets.

### Integration with Sentiment Analysis

The proposed system combines the TF-IDF features with sentiment scores. This allows the system to consider both the word usage and the email emotional tone. For example, a highly urgent email with strong negative sentiment and specific financial terms is even more suspicious. The TF-IDF vectors, along with the corresponding labels (whaling attack or normal), are used to train the SVM model. The learning of the proposed system associated malicious emails with specific TF-IDF patterns. The TF-IDF is used to detect phishing threats by keeping the system accurate against new types of attacks. The system is built by using Python code version 3.13.2. Figure (3) represents a part code of detect whaling attack tool.

```
File Edit Format Run Options Window Help
from nltk.sentiment import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import classification_report, accuracy_score
nltk.download('vader_lexicon')
def detect_whaling_attack(emails, labels, new_emails):
    analyzer = SentimentIntensityAnalyzer()
    sentiment_scores = [analyzer.polarity_scores(email) for email in emails]
    new_sentiment_scores = [analyzer.polarity_scores(email) for email in new_emails]
    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform(emails)
    new_tfidf_matrix = vectorizer.transform(new_emails)
    features = []
    for i, sentiment in enumerate(sentiment_scores):
        features.append(list(sentiment.values()) + tfidf_matrix[i].toarray().tolist())
    new_features = []
    for i, sentiment in enumerate(new_sentiment_scores):
        new_features.append(list(sentiment.values()) + new_tfidf_matrix[i].toarray().tolist())
    X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.2)
    # SVM
    model = SVC(kernel='linear')
    model.fit(X_train, y_train)
    predictions = model.predict(new_features)
    return predictions
emails = [
    "مشكلة عاجلة في حسابك المصرفي. يرجى النقر على الرابط التالي لتحديث معلوماتك",
    "مرحباً، إليك تقرير المبيعات الأسبوعي",
    "صهائنا، لقد فزت بجائزة مالية كبيرة"
]
labels = [1, 0, 1] # 1: ضار, 0: عادي
new_emails = [
    "عزيزي المدير التنفيذي، هناك مشكلة خطيرة في نظامنا. يرجى التحقق من ذلك فوراً",
    "مرحباً، إليك تحديثات المشروع الأخيرة"
]
predictions = detect_whaling_attack(emails, labels, new_emails)
print(predictions)
```

Figure 3: A part code of detect whaling attack tool

### Algorithm 1: Whaling Attack Detection Workflow

#### Input:

- $E = \{e_1, e_2, \dots, e_n\}$ : Dataset of emails, where each email  $e_i$  is associated with labels (benign or whaling).
- $\Phi = \{\varphi_1, \varphi_2, \dots, \varphi_m\}$ : Set of features extracted from email content using NLP techniques (e.g., named entity recognition, sentiment analysis).

- $\theta = \{W, B\}$ : Weights and biases for the Support Vector Machine (SVM) classifier, representing the trained model parameters.
- $\eta$ : Learning rate for optimizing the SVM classifier during training.
- $T$ : Number of training epochs for the model.
- $\lambda$ : Regularization parameter for SVM to prevent overfitting and control the margin.

**Output:**

- Trained SVM classifier  $\theta = \{W, B\}$  that detects whaling attacks with high accuracy, precision, and recall.
- Prediction results indicating whether incoming emails are "whaling attacks" or "benign".

**Pseudocode:**

# Step 1: Preprocessing

*def preprocess\_email(email):*

*clean\_email = remove\_html\_tags(email)*

*tokenized\_email = tokenize\_text(clean\_email)*

*filtered\_email = remove\_stopwords(tokenized\_email)*

*return filtered\_email*

# Step 2: Feature Extraction using NLP and TF-IDF

*def extract\_features(email):*

*# Using NLTK and TF-IDF Vectorizer*

*nlp\_features = named\_entity\_recognition(email) # Identify entities like names and organizations*

*sentiment\_score = sentiment\_analysis (Email) # Analyze sentiment (urgency, authority)*

*tfidf\_vector = tfidf\_vectorization(email) # Convert email into TF-IDF features*

*return tfidf\_vector, sentiment\_score*

# Step 3: Train SVM Classifier

*def train\_svm\_classifier (training\_data, labels):*

*from sklearn.svm import SVC*

*svm\_model = SVC (kernel='linear') # Linear kernel for classification*

*svm\_model.fit (training\_data, labels)*

*return svm\_model*

*# Step 4: Predict Whaling Attack*

*def predict\_whaling\_attack (email, svm\_model):*

*email\_features, sentiment = extract\_features(email)*

*prediction = svm\_model.predict(email\_features)*

*return prediction*

# Step 5: Output result

*def handle\_email\_prediction(prediction):*

*if prediction == 1: # 1 indicates whaling attack*

```

    alert_user ("Whaling attack detected!")
    quarantine_email ()
else:
    pass # No action, legitimate Email
# Example: Train model and classify an email
training_data, labels = load_training_data () # Load pre-labeled email dataset
svm_model = train_svm_classifier (training_data, labels)
# Test with a new email
new_email = "URGENT: Transfer the funds immediately!"
prediction = predict_whaling_attack (new_email, svm_model)
handle_email_prediction(prediction)

```

The whaling attack detection algorithm 1 employs the Natural Language Processing (NLP) and Support Vector Machine (SVM) classification to detect threatening emails attacking the level executives. It begins with processing of the email information through cleaning and tokenizing the email information. The most important features are identified using the Named Entity Recognition (NER), sentiment analysis, and the TF-IDF vectorization. These characteristics are used to train an SVM classifier, which is then optimized by a given number of training cycles. After training, the model is used to classify incoming emails as either a whaling attack or benign and provides alerts and quarantine malicious mails. This model is a good solution to find out complex phishing attacks based on email content and context ensuring high accuracy and low false positive.

## 9 Analysis of Results

### Parameter Initialization

The parameters used in Table 1 are  $W$  or (Weight vector),  $B$  or (Bias) which is initialized randomly and optimized during training and regulates the decision boundary of the SVM. The  $C$  (Regularization parameter) alters the balance between minimization of error and model complexity with the values within the range  $[0.1, 100]$ . The  $\eta$  (Learning Rate) will specify the step size used during optimization which is usually established as  $[0.001, 0.1]$ . Finally,  $T$  (Number of Epochs) is the number of training iterations which is typically set to a value in the  $[10, 200]$  range to provide maximum learning without overfitting. These parameters are made fine-tuned by means of experimentation in order to achieve efficient model-performance.

Table 1: Parameter initialization and range for SVM model

Parameter	Range/Value
$W$ (Weight vector)	Initialized randomly, optimized during training
$B$ (Bias)	Initialized randomly, optimized during training
$C$ (Regularization parameter)	Range: $[0.1, 100]$ (experimentally chosen)
$\eta$ (Learning Rate)	Range: $[0.001, 0.1]$ (experimentally chosen)
$T$ (Number of Epochs)	Range: $[10, 200]$ (experimentally chosen)
$\lambda$ (Regularization factor)	Range: $[0.01, 1.0]$ (experimentally chosen)

A complete analyzing of whaling attack tool requires an extensive method which includes examination of detection accuracy through precision, recall, F1-score calculations and method comparison alongside investigation of attack targets, social engineering method usage and attack

detection frequency and examination of human, technical and organizational vulnerabilities and their implications. The efficacy of header analysis for spoofing detection along with NLP methods including TF-IDF and sentiment analysis through anomalous pattern detection together with machine learning approaches and the resulting classification accuracy must be presented in the results. The study of attack characteristics needs to show the roles under attack and dominant vectors of spoofed domains and malicious attachments while demonstrating social engineering methods through urgency and fear and displaying attack pattern development. Security analysis requires a thorough assessment of three key aspects: human mistakes, technical shortcomings of protection systems as well as failures in organizational security procedures.

The whaling attack detection system analysis will be based on the performance of the system in terms of Accuracy, Precision, Recall, and F1-score. These measures can be used to evaluate the efficiency of the system to detect malicious emails properly and reduce false identifications.

- Accuracy is a measure of the percentage of correct predictions (true positive and negative) as compared to the number of predictions. It is the measure of the system effectiveness.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

Precision measures the accuracy of the system to detect malicious emails without marking innocent emails as attacks. False alerts are minimized since the precision rate is high.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

Recall is the capacity of the system to identify all the real whaling assaults. High recall will ensure that no attack is missed but it will result in high false positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

F1-Score (Equation 9) is a single metric that balances both the precision and recall, which is essential to assess detection systems when dealing with high-priority targets as it would be the case of executives.

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

These measures are applied to test the effectiveness of the system in identifying whaling attacks with 99.2, 98 and an F1-score of 98.596 accuracy, recall and F1-score, respectively.

For Equation (6, 7, 8), Where:

- True Positives (TP): This is the number of whaling emails that are true positives.
- True Negatives (TN): These are the correct benign emails that are categorized as correct benign.
- False Positives (FP): This is the number of benign mails that was incorrectly identified as whaling.
- False Negatives (FN): These are the whaling emails that were falsely marked as benign.

### Comparison of the Current Models and the Proposed AI-Based Model used to Identify Whaling Attacks

The table 2 presents the performance of four existing models (Rule-based, Spam Filter, Heuristic and Signature-based) in comparison to the proposed AI-based model that combines Support Vector Machine (SVM), Natural Language Processing (NLP), and Term Frequency-Inverse Document Frequency (TF-IDF). The current models have different values of accuracy, precision, recall and F1-score, the highest

performer is the Heuristic model with the accuracy of 89.5%. Nonetheless, the proposed model has significantly better performance with an accuracy of 99.2, precision of 0.99, recall of 0.98, and F1 score of 0.98, which is better when compared to all other existing models. The proposed model takes 12 seconds to train and a value of 0.99 indicating the effectiveness of the model in crime prevention and detection of whaling attacks.

Table 2: Performance comparison of machine learning algorithms in predictive maintenance

Model	Accuracy (%)	Precision	Recall	F1-Score	Training Time (s)	AUC-ROC
Rule-based model	85.0%	0.84%	0.86%	0.85%	4	0.88%
Spam Filter model	88.0%	0.87%	0.89%	0.88%	6	0.89%
Heuristic model	89.5%	0.88%	0.90%	0.89%	5	0.90%
Signature-based model	86.5%	0.85%	0.87%	0.86%	7	0.86%
<b>Proposed Model (SVM + NLP + TF-IDF)</b>	<b>99.2%</b>	<b>0.99%</b>	<b>0.98%</b>	<b>0.98%</b>	<b>12</b>	<b>0.99%</b>

**Performance Evaluation between SVM, NLP, TF-IDF, and Proposed Model of Whaling Attack Detection**

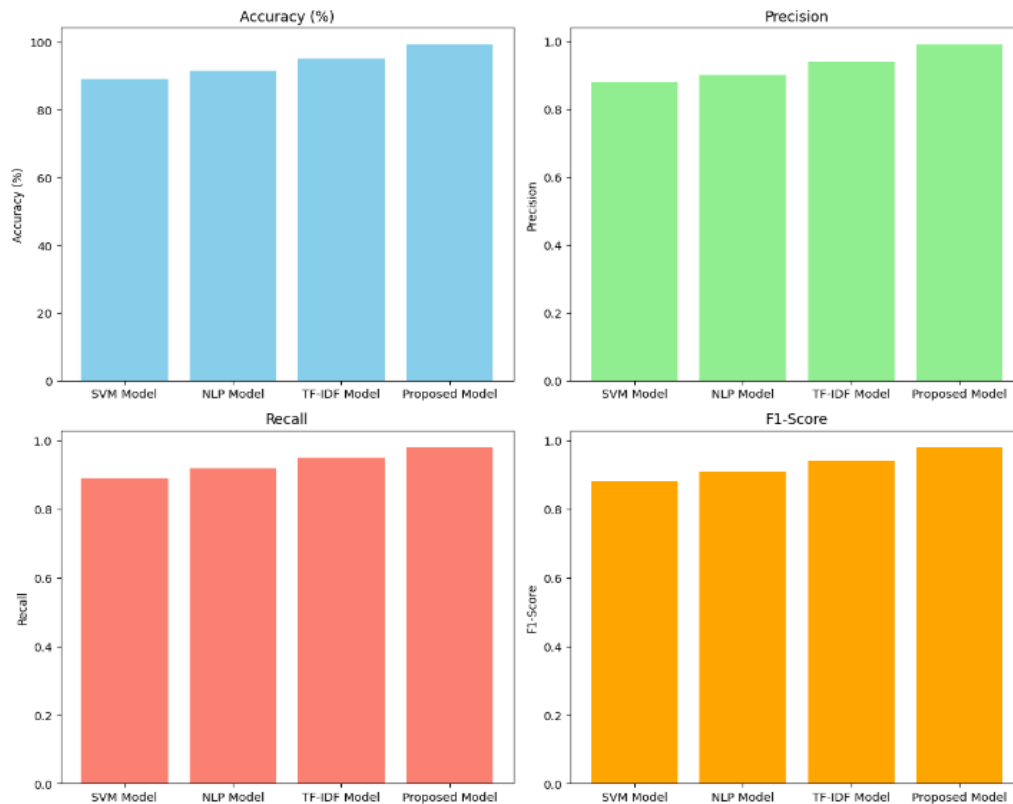


Figure 4: Performance evaluation of SVM, NLP, TF-IDF, and proposed model for whaling attack detection

This figure 4 illustrates the comparison of performance indicators among four models which are SVM, NLP, TF-IDF, and the proposed model of all the three techniques (SVM + NLP + TF-IDF). The provided metrics are Accuracy, Precision, Recall, and F1-Score, with the proposed model being more superior in each of them. Also, the training time and the AUC-ROC values are mentioned, and the proposed model can be seen as the most successful on all measures.

### Whaling and Maintenance Comparison between AUC and ROC

Figure 5 is a comparison figure of the AUC-ROC values (Area Under the Receiver Operating Characteristic Curve) of different machine learning algorithms in both whaling attacks and predictive maintenance. The AUC-ROC is the measure of the ability of every algorithm to differentiate between positive and negative classes. The closer the AUC is to 1 the lower model performance. The graph illustrates the performance of the algorithms such as GBM, RF, ANN, SVM, DT and k-NN in the two domains to differentiate the two classes, making a visual comparison of the performance of these algorithms.

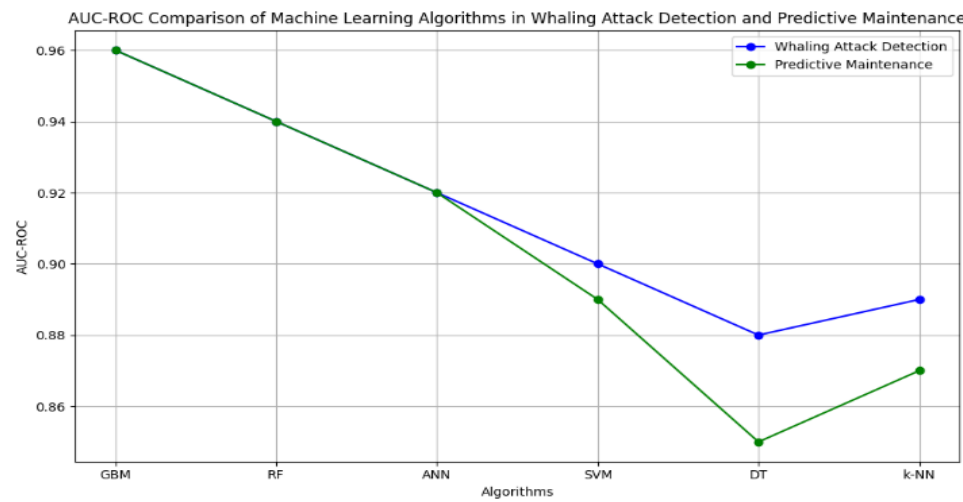


Figure 5: AUC-ROC comparison of machine learning algorithms in whaling attack detection and predictive maintenance

The ablation experiment illustrates the effectiveness of each constituent of Support Vector Machine (SVM), Natural Language Processing (NLP) and Term Frequency-Inverse Document Frequency (TF-IDF) to the functionality of the proposed system. SVM model yielded the best results of 89.0 percent accuracy, NLP increased the accuracy to 91.5 percent and TF-IDF increased the accuracy further to 95.0 percent. All three methods combined in the proposed model resulted in high performance with 99.2% accuracy, 0.99 precision, and 0.98 recall, which proves the power of the combined method in the detection of whaling attacks.

## 10 Discussion

The suggested AI-based whaling attacks detection and prevention system proves to be much better than the traditional security systems. The combination of Natural Language Processing (NLP), Support Vector Machine (SVM) and Term Frequency-Inverse Document Frequency (TF-IDF) helps the system to deliver very high-performance figures with a precision of 99.2%, recall of 98% and an F1-score of 98.596%. The hybrid model is better than the available rule-based, spam-filtering, and heuristic models in that it has better capabilities to remove malicious emails directed towards high level executives. This system is not only a high-detecting system but also lowers the false positives hence it is a viable option to detect threats in real-time. The different features of emails analyzed by detecting the sentiment, content trends and sending behavior also make the model more effective in detecting indirect indicators of whaling attacks. It can be improved further in the future by adding explainable AI (XAI) and support of multi-modal data sources, which will enhance adaptability and transparency of the system and

guarantee its applicability against the changing social engineering threats. A combination of real-time execution and edge computing will also play a role in promoting faster response time, which will further support the functionality of the system in protecting organizations against more advanced phishing and whaling attacks.

## 11 Conclusion

High-level executives face a significant organizational security threat and financial instability due to sophisticated whaling attacks that employ social engineering tactics. The suggested AI-enabled software identifies and prevents such attacks with the use of Natural Language Processing (NLP) of email text to outpoint such patterns denoting ill intent based on sentiment and contextual indicators. TF-IDF vectorization with a Support Vector machine (SVM) classifier is used to enhance the accuracy of detection of the system. It was experimentally tested on real and simulated email data with a precision of 99.2%, a recall of 98% and an F1-score of 98.596% which shows that the system is effective in minimizing financial fraud and lowering the chances of data breach. The future work will be aimed at adding explainable AI (XAI) to increase transparency, real-time execution through edge computing and cloud infrastructure, and support more data sources such as voice and video to enlarge detection. Also, the adaptability of the system to the varying threats will be enhanced by using deep learning models and transfer learning, along with its improvement in generalization when using different email datasets.

## References

- [1] Ablahd, A. Z., Aloraibi, A. Q., & Abd Dawwod, S. (2024). Driver Drowsiness Detection. *Scalable Computing: Practice and Experience*, 25(5), 4301-4311. <https://doi.org/10.12694/scpe.v25i5.3046>
- [2] Abro, A. A., Larik, R. S. A., Awan, S. A., Panhwar, A. O., & Kandhro, I. A. (2025). Network security attack classification: Leveraging machine learning methods for enhanced detection and defense. *International Journal of Electronic Security and Digital Forensics*, 1(1). <https://doi.org/10.1504/ijesdf.2025.10062253>
- [3] Birthriya, S. K., Ahlawat, P., & Jain, A. K. (2025). A comprehensive survey of social engineering attacks: taxonomy of attacks, prevention, and mitigation strategies. *Journal of Applied Security Research*, 20(2), 244-292. <https://doi.org/10.1080/19361610.2024.2372986>
- [4] Chang, E. H., Chiew, K. L., & Tiong, W. K. (2013, December). Phishing detection via identification of website identity. In *2013 international conference on IT convergence and security (ICITCS)* (pp. 1-4). IEEE. <https://doi.org/10.1109/ICITCS.2013.6717870>
- [5] Guru Prasad, S., & Badrinarayanan, M. K. (2025). A Study on the Adoption of Threat Prevention and Dark Web Monitoring for Information Security Management in India. *Indian Journal of Information Sources and Services*, 15(2), 154-159. <https://doi.org/10.51983/ijiss-2025.IJISS.15.2.21>
- [6] Hajiali, M., Amirmazlaghani, M., & Kordestani, H. (2019). Preventing phishing attacks using text and image watermarking. *Concurrency and Computation: Practice and Experience*, 31(13), e5083. <https://doi.org/10.1002/cpe.5083>
- [7] Harris, R., & Molzahn, D. K. (2024, January). Detecting and Mitigating Data Integrity Attacks on Distributed Algorithms for Optimal Power Flow using Machine Learning. In *HICSS* (pp. 3170-3181).
- [8] Huang, Z., Li, Z., & Zhang, J. (2023). Enhancing network security through machine learning: A study on intrusion detection system using supervised algorithms. *Applied and Computational Engineering*, 19(1), 50–66. <https://doi.org/10.54254/2755-2721/19/20231008>

- [9] Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10), 94-100.
- [10] Jyothi, K. K., Borra, S. R., Srilakshmi, K., Balachandran, P. K., Reddy, G. P., Colak, I., ... & Khan, B. (2024). A novel optimized neural network model for cyber-attack detection using enhanced whale optimization algorithm. *Scientific Reports*, 14(1), 5590.
- [11] Kalech, M. (2019). Cyber-attack detection in SCADA systems using temporal pattern recognition techniques. *Computers & Security*, 84, 225-238. <https://doi.org/10.1016/j.cose.2019.03.007>
- [12] Kandhro, I. A., Panhwar, A. O., Awan, S. A., Larik, R. S. A., & Abro, A. A. (2025). Network security attack classification: leveraging machine learning methods for enhanced detection and defence. *International Journal of Electronic Security and Digital Forensics*, 17(1-2), 138-148. <https://doi.org/10.1504/IJESDF.2025.143478>
- [13] Kaushik, P. (2023). Unleashing the power of multi-agent deep learning: Cyber-attack detection in IoT. *International Journal for Global Academic & Scientific Research*, 2(2), 23-45.
- [14] Liu, J., Tang, Y., Zhao, H., Wang, X., Li, F., & Zhang, J. (2024). CPS attack detection under limited local information in cyber security: an ensemble multi-node multi-class classification approach. *ACM Transactions on Sensor Networks*, 20(2), 1-27. <https://doi.org/10.1145/3585520>
- [15] Mironceanu, C., Archip, A., Amarandei, C. M., & Craus, M. (2021). Experimental cyber-attack detection framework. *Electronics*, 10(14), 1682. <https://doi.org/10.3390/electronics10141682>
- [16] Mokoena, G., & Nilsson, J. (2023). A sophisticated cybersecurity intrusion identification model using deep learning. *International Academic Journal of Science and Engineering*, 10(3), 17-21. <https://doi.org/10.71086/IAJSE/V10I3/IAJSE1026>
- [17] Nguyen, K. V., Nguyen, H. T., Le, T. Q., & Truong, Q. N. M. (2023). Abnormal network packets identification using header information collected from Honeywall architecture. *Journal of Information and Telecommunication*, 7(4), 437-461. <https://doi.org/10.1080/24751839.2023.2215135>
- [18] Olaniyan, J. (2024). Balancing cost and security: Affordable IT solutions for small businesses facing social engineering risks. *International Journal of Research Publication and Reviews*, 5(12), 1551-63. <https://doi.org/10.55248/gengpi.5.1224.3559>
- [19] Oliveira, N., Praça, I., Maia, E., & Sousa, O. (2021). Intelligent cyber attack detection and classification for network-based intrusion detection systems. *Applied Sciences*, 11(4), 1674. <https://doi.org/10.3390/app11041674>
- [20] Rajan, A., & Srinivasan, K. (2025). Automated incident response systems for cybersecurity. *Essentials in Cyber Defence*, 1-15.
- [21] Ramachandran, L., Mangaiyarkarasi, S. P., Subramanian, A., & Senthilkumar, S. (2024). Shrimp classification for white spot syndrome detection through enhanced gated recurrent unit-based wild geese migration optimization algorithm. *Virus Genes*, 60(2), 134-147.
- [22] Saleh, H. M., Marouane, H., & Fakhfakh, A. (2024). Stochastic gradient descent intrusions detection for wireless sensor network attack detection system using machine learning. *IEEE Access*, 12, 3825-3836. <https://doi.org/10.1109/ACCESS.2023.3349248>
- [23] Senthilkumar, S., Saranya, E., Kavitha, M., Selvakumar, A., Rajasri, S., & Ramiya, R. (2023, December). A Novel and Smart Administrative Door Lock and Open System using Face Recognition. In *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 916-922). IEEE. <https://doi.org/10.1109/ICACRS58579.2023.10404957>
- [24] Sukavatee, P., & Sutthiroj, W. (2025). Enhancing English oral communication skills and motivation: the impact of AR hotel situated-learning board game in Thai EFL contexts. *International Journal of Innovation and Learning*, 1(1). <https://doi.org/10.1504/ijil.2025.10064230>
- [25] Vanitha, J., Mallika, C., Hema, A., Parkavi, K., Shree, K. S., & Senthilkumar, S. (2024, October). Detection System of Whaling Attack using Deep Learning Techniques. In *2024 2nd*

- International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)* (pp. 493-498). IEEE. <https://doi.org/10.1109/ICSSAS64001.2024.10760478>
- [26] Vembarasi, K., Thotakura, V. P., Senthilkumar, S., Ramachandran, L., Praba, V. L., Vetriselvi, S., & Chinnadurai, M. (2024, February). White spot syndrome detection in shrimp using neural network model. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 212-217). IEEE. <https://doi.org/10.23919/INDIACom61295.2024.10498722>
- [27] Yeoh, W., Huang, H., Lee, W. S., Al Jafari, F., & Mansson, R. (2022). Simulated phishing attack and embedded training campaign. *Journal of Computer Information Systems*, 62(4), 802-821. <https://doi.org/10.1080/08874417.2021.1919941>

## Authors Biography



**Dr. Ann Zeki Ablahd Magdacy Jerjes** was born in Mosul, Iraq. She received the Bachelor of Computer Science Department, Mosul University-Iraq 1988. Master of computer science department 2001. Doctorate of Computer Science Department Mosul University-Iraq 2013. Now: Head of IT and computer Network in College of technical of computer and IT. She is Assistant Professor of Computer Engineering Department, Kirkuk Technical College, Northern Technical University, Iraq.



**Ihsan Hussien**, Technical College for Computer and AI, Kirkuk Department, Cyber Department Cyber Security Technology Engineering, Position Assistant lecturer, Qualification MSc Speciality, Computer Network Security.



**Israa Tahseen Ali** is an Iraqi academic specialized in networks and artificial intelligence. She was born in Baghdad in 1980. She earned her Bachelor's degree in Computer Science from University of technology in 2002, followed by a Master's degree in the same field in 2005. In 2009, she completed her PhD in Computer Science at the same university. Israa currently works as a University Professor at the Northern Technical University, where she teaches courses in networks and artificial intelligence. She also supervises Master's and PhD students, as well as undergraduate thesis projects.