AI-BASED PHISHING DETECTION WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

Dr. M. Navaneetha Krishnan,
Professor and Head of Department, Computer Science and Engineering,
St. Joseph College of Engineering,Chennai-602117, Tamil Nadu,
Email Id – mnksjce@gmail.com

Ms. Ajasha Alex,
Student, Computer Science and Engineering,
St. Joseph College of Engineering ,Chennai-602117, Tamil Nadu,
Email Id – ajashaalex01@gmail.com

Ms. Dishney Classin,
Student, Computer Science and Engineering,
St. Joseph College of Engineering, Chennai-602117, Tamil Nadu,
Email Id – dishney70@gmail.com

ABSTRACT

Phishing attacks are one of the most common cybersecurity threats that attempt to steal sensitive information such as passwords, banking details, and personal data by impersonating trusted sources. Traditional phishing detection systems rely on rule-based methods and basic machine learning techniques, which often fail to detect advanced and newly emerging phishing attacks. Moreover, these systems provide only binary results without explaining the reason behind their predictions, reducing user trust and understanding.

This project proposes an AI-Based Phishing Detection System with Explainable AI (XAI) that uses machine learning and Natural Language Processing (NLP) techniques to analyse emails, URLs, and messages. The system extracts important features such as suspicious keywords, URL patterns, and domain information to accurately classify inputs as phishing or

safe. In addition to detection, the system provides a risk score and clear, human-readable explanations using XAI techniques, helping users understand why a particular input is flagged.

The proposed system improves detection accuracy, enhances transparency, and increases user awareness. By integrating explainability and continuous learning, the system provides a reliable and effective solution for preventing phishing attacks and strengthening cybersecurity.

KEYWORDS

Phishing Detection, Explainable AI (XAI), Machine Learning, Natural Language Processing (NLP), Cybersecurity, Feature Extraction, Risk Scoring, Email and URL Analysis, Classification, Threat Detection

I. INTRODUCTION

With the rapid growth of internet usage, cybersecurity threats have increased significantly, and phishing has become one of the most common and dangerous attacks. Phishing is a type of cyberattack in which attackers impersonate trusted organizations or individuals to trick users into revealing sensitive information such as usernames, passwords, banking details, and personal data. These attacks are usually carried out through emails, fake websites, and malicious links that appear legitimate, making them difficult for users to identify. Traditional phishing detection methods rely on blacklist databases, rule-based systems, and basic machine learning techniques. While these methods can detect known phishing attacks, they often fail to identify new and sophisticated phishing attempts. In addition, most existing systems provide only a binary result (phishing or safe) without explaining the reason behind the decision. This lack of transparency reduces user trust and makes it difficult for users to understand and avoid such threats in the future.

To overcome these limitations, Artificial Intelligence (AI) and Machine Learning (ML) techniques have been widely used to improve phishing

detection accuracy. These techniques analyse patterns, extract features from emails and URLs, and identify suspicious characteristics. However, many AI-based systems act as “black boxes,” meaning users cannot understand how the system made its decision. This creates a need for Explainable AI (XAI), which provides clear and understandable explanations for predictions. This project proposes an AI-Based Phishing Detection System with Explainable AI that uses machine learning and Natural Language Processing (NLP) techniques to analyse emails, URLs, and messages. The system extracts relevant features, classifies the input, generates a risk score, and provides human-readable explanations for the prediction. This approach not only improves detection accuracy but also increases transparency, user trust, and cybersecurity awareness. The proposed system can be integrated with real-world applications to provide effective and reliable protection against phishing attacks.

II. BACKGROUND AND MOTIVATION

The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness. The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness. The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness. The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect

dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness. The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness.

The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness. The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness. The evolution of phishing attacks has increased the need for intelligent detection mechanisms. Static signature-based methods cannot detect dynamically generated phishing domains. Machine learning models such as Random Forest and SVM have demonstrated improved performance. Explainability ensures regulatory compliance, auditing transparency, and improved user awareness.

III. SYSTEM ARCHITECTURE

The system architecture of the AI-Based Phishing Detection with Explainability consists of multiple modules that work together to detect phishing attacks and provide clear explanations. The system starts by receiving input from the user in the form of emails, URLs, or text messages.

The input is then processed, analysed, and classified using machine learning techniques, and the result is displayed along with an explanation.

Components of System Architecture

1. Input Layer

This layer accepts data from the user, such as emails, URLs, or messages. It acts as the entry point of the system and sends the input to the preprocessing module.

2. Preprocessing Layer

This layer cleans and prepares the input data by removing unwanted symbols, converting text to lowercase, and tokenizing the text. It ensures the data is in the proper format for analysis.

3. Feature Extraction Layer

This layer extracts important features from the input, such as suspicious keywords, URL length, domain information, and special characters. These features help identify phishing patterns.

4. Detection Layer (AI/ML Model)

This layer uses machine learning algorithms such as Random Forest, Decision Tree, or SVM to classify the input as phishing or legitimate. It also generates a risk score based on the analysis.

5. Explainability Layer (XAI Module)

This layer uses Explainable AI techniques such as LIME or SHAP to provide clear explanations for the prediction. It highlights the important features that influenced the decision.

6. Output Layer

This layer displays the final result to the user, including the classification (phishing or safe), risk score, and explanation. This helps users understand the threat and take appropriate action.

Architecture Flow

Step 1: User Input

The process begins when the user provides an input such as an email, URL, or text message. This input is collected by the system through a user interface or application.

Step 2: Data Preprocessing

The input data is cleaned and prepared for analysis. This includes converting text to lowercase, removing special characters and stop words, and tokenizing the text. This step ensures the data is in a suitable format for feature extraction.

Step 3: Feature Extraction

In this step, important phishing-related features are extracted from the input. These features include suspicious keywords, URL length, domain information, number of special characters, and other patterns that indicate phishing.

Step 4: Phishing Detection using AI/ML Model

The extracted features are given to a trained machine learning model such as Random Forest, Decision Tree, or Support Vector Machine. The model analyses the features and classifies the input as phishing or legitimate. It also generates a risk score to indicate the severity of the threat.

Step 5: Explainability using XAI Module

The Explainable AI module analyses the model's prediction and identifies the key features that influenced the decision. It provides a clear and human-readable explanation, helping users understand why the input is classified as phishing or safe.

Step 6: Output Generation

Finally, the system displays the result to the user. The output includes the classification result (phishing or safe), the risk score, and the explanation. This helps the user understand the threat and take appropriate action.

IV. Module Description

The proposed system consists of several modules that work together to detect phishing attacks and provide explainable results. Each module performs a specific function in the phishing detection process.

1. Input Module

This module collects input data from the user, such as emails, URLs, or text messages. It acts as the entry point of the system and ensures the input is properly received for further processing. The collected data is then passed to the preprocessing module.

2. Preprocessing Module

This module cleans and prepares the input data for analysis. It removes unnecessary symbols, punctuation, and stop words, and converts text into lowercase. It also tokenizes the text into meaningful words. This step improves the quality of data and ensures accurate feature extraction.

3. Feature Extraction Module

This module extracts important features from the input data that help identify phishing attacks. These features include URL length, presence of special characters, suspicious keywords, domain information, and other phishing indicators. The extracted features are converted into a structured format for the detection module.

4. Detection Module (AI/ML Model)

This module uses machine learning algorithms such as Random Forest, Decision Tree, or Support Vector Machine to analyse the extracted features. It classifies the input as phishing or legitimate and generates a risk score based on the level of threat detected.

5. Explainability Module (XAI Module)

This module provides clear and understandable explanations for the model's prediction using Explainable AI techniques such as LIME or SHAP. It highlights the key features that influenced the decision, helping users understand why the input is classified as phishing or safe.

6. Output Module

This module displays the final result to the user. It shows the classification result (phishing or safe), the risk score, and the explanation. This helps users understand the threat and take appropriate action.

V. IMPLEMENTATION METHODOLOGY

The implementation methodology describes the step-by-step process used to develop the AI-Based Phishing Detection System with Explainability. The system is implemented using machine learning, natural language processing, and explainable AI techniques to detect phishing attacks and provide clear explanations.

Step 1: Data Collection

Phishing and legitimate datasets are collected from reliable sources such as Phish Tank and Kaggle. The dataset includes emails, URLs, and text messages labelled as phishing or legitimate. This data is used for training and testing the machine learning model.

Step 2: Data Preprocessing

The collected data is cleaned and prepared for analysis. This includes removing unnecessary symbols, converting text to lowercase, removing stop words, and tokenizing the text. This step ensures the data is consistent and suitable for feature extraction.

Step 3: Feature Extraction

Important features are extracted from the pre-processed data. These features include URL length, presence of special characters, suspicious keywords, domain information, and other phishing indicators. The extracted features are converted into numerical format for model training.

Step 4: Model Training

Machine learning algorithms such as Random Forest, Decision Tree, or Support Vector Machine are used to train the model. The model learns patterns from the dataset and identifies differences between phishing and legitimate inputs.

Step 5: Model Testing and Evaluation

The trained model is tested using test data to evaluate its performance. Metrics such as accuracy, precision, recall, and F1-score are used to measure the effectiveness of the model.

Step 6: Explainable AI Integration

Explainable AI techniques such as LIME or SHAP are integrated with the trained model. These techniques provide explanations by highlighting the important features that influenced the prediction.

Step 7: System Integration

All modules, including input, preprocessing, feature extraction, detection, explainability, and output, are integrated into a complete system. This allows users to provide input and receive detection results with explanations.

Step 8: Output Generation

The system displays the final result, including classification (phishing or safe), risk score, and explanation. This helps users understand the threat and take appropriate action.

VI. RECENT ADVANCEMENTS IN FEDERATED LEARNING AND PRIVACY

Federated Learning (FL) has emerged as a powerful approach to train machine learning models while preserving user privacy. Instead of sending

raw data to a central server, FL allows models to be trained locally on user devices and only shares model updates. This approach is especially useful in phishing detection systems where email and user data are sensitive.

1. Privacy-Preserving Model Training

Federated Learning ensures that sensitive data such as emails, URLs, and user credentials remain on local devices. Only model parameters like weights and gradients are shared with the central server. This significantly reduces the risk of data leakage and unauthorized access.

Key Benefits:

- Protects confidential user information
- Reduces risk of centralized data breaches
- Complies with privacy regulations

2. Secure Aggregation Techniques

Recent advancements include secure aggregation methods that combine model updates from multiple devices without revealing individual contributions. Encryption techniques such as homomorphic encryption and secure multi-party computation ensure that even the server cannot see individual data patterns.

Impact:

- Prevents exposure of individual device information
- Enhances trust in distributed learning systems
- Improves security in phishing detection models

3. Differential Privacy Integration

Differential privacy adds controlled noise to model updates before sharing them with the server. This ensures that individual user data cannot be reverse-engineered from the trained model.

Advantages:

- Provides mathematical privacy guarantees
- Prevents reconstruction of sensitive data
- Improves user trust and system reliability

4. Edge-Based Intelligence for Real-Time Detection

Modern systems use edge computing with federated learning, allowing phishing detection models to run directly on user devices such as smartphones and laptops. This enables faster detection without sending data to external servers.

Benefits:

- Faster phishing detection
- Reduced latency
- Improved user privacy

5. Robustness Against Adversarial Attacks

Advanced federated learning systems include mechanisms to detect malicious participants that attempt to poison the model. Techniques like anomaly detection and trust scoring help ensure model integrity.

Outcome:

- Improved model reliability
- Protection against model poisoning
- Enhanced system robustness

6. Personalized Federated Learning Models

Recent research focuses on creating personalized models for individual users while still benefiting from global knowledge. This improves phishing detection accuracy based on user-specific patterns.

Advantages:

- Higher detection accuracy
- Better adaptation to user behaviour
- Improved performance in diverse environments

VII. ADVANCEMENTS IN PHISHING DETECTION TECHNOLOGIES

Phishing detection technologies have evolved significantly from traditional rule-based systems to advanced Artificial Intelligence (AI) and Machine Learning (ML) approaches. Modern systems focus on improving detection accuracy, reducing false positives, and providing real-time protection against sophisticated phishing attacks.

1. Machine Learning-Based Detection

Machine Learning algorithms have improved phishing detection by automatically learning patterns from large datasets. Unlike traditional blacklist methods, ML models can detect previously unseen phishing attacks.

Key Features:

- Uses classifiers such as Decision Tree, Random Forest, and Support Vector Machine (SVM)
- Detects phishing based on URL structure, domain features, and webpage content
- Improves detection accuracy over time

Impact:

- Detects zero-day phishing attacks
- Reduces dependence on manual rule creation

2. Deep Learning Approaches

Deep Learning models provide higher accuracy by automatically extracting complex features from phishing data.

Common Models Used:

- Convolutional Neural Networks (CNN) for webpage analysis
- Recurrent Neural Networks (RNN) and LSTM for sequence analysis
- Neural Networks for email content classification

Advantages:

- Automatically learns hidden patterns
- Provides better detection accuracy than traditional ML

3. Natural Language Processing (NLP) for Email Analysis

NLP techniques analyse email content to detect phishing attempts based on language patterns.

Capabilities:

- Detects urgency words (e.g., “urgent”, “verify now”)
- Identifies brand impersonation
- Analyses semantic meaning of email text

Benefits:

- Improves detection of email-based phishing
- Detects socially engineered attacks

4. URL and Website Behaviour Analysis

Advanced systems analyse URL structure and website behaviour to detect suspicious activities.

Features Analysed:

- Domain age and registration details

- Presence of special characters or suspicious patterns
- Website loading behaviour and redirects

Outcome:

- Early detection of malicious websites
- Improved accuracy in identifying fake domains

5. Real-Time Phishing Detection Systems

Modern phishing detection systems operate in real time to protect users instantly.

Technologies Used:

- Browser extensions
- Email security gateways
- Cloud-based detection systems

Benefits:

- Immediate threat detection
- Prevents users from accessing phishing sites

6. Explainable AI (XAI) for Transparent Detection

Explainable AI provides clear explanations for phishing detection decisions.

Advantages:

- Improves user trust
- Helps security experts understand detection reasons
- Supports better decision-making

Example Explanations:

- Suspicious URL structure

- Presence of phishing keywords
- Domain reputation issues

7. Federated Learning for Privacy-Preserving Detection

Federated Learning allows phishing detection models to be trained without sharing sensitive user data.

Benefits:

- Protects user privacy
- Enables secure distributed learning
- Improves detection performance

8. Integration with Threat Intelligence Systems

Modern systems integrate global threat intelligence feeds to improve detection.

Capabilities:

- Uses updated phishing databases
- Shares threat information across systems
- Enhances overall security

VIII. CHALLENGES

Phishing detection systems face several challenges due to the evolving nature of cyberattacks and the increasing sophistication of phishing techniques.

1. Rapid Evolution of Phishing Attacks

Phishers continuously change their techniques to bypass detection systems. They use new domains, URL obfuscation, and advanced social engineering methods, making detection more difficult.

2. Detection of Zero-Day Attacks

Zero-day phishing attacks are new attacks that are not present in existing blacklists or databases. Traditional systems fail to detect these unknown threats effectively.

3. High False Positive Rate

Some legitimate websites or emails may be incorrectly classified as phishing. This reduces user trust and affects the usability of the system.

4. Lack of Explainability in AI Models

Many AI and Deep Learning models act as black boxes, making it difficult to understand why a prediction was made. This reduces transparency and trust in the system.

5. Handling Large and Dynamic Data

Phishing detection systems must process large volumes of emails, URLs, and web data in real time. Managing and analysing this data efficiently is challenging.

6. Privacy and Security Concerns

Collecting and analysing user data may create privacy risks. Protecting sensitive user information while performing detection is an important challenge.

7. Adversarial Attacks on AI Models

Attackers may manipulate input data to fool machine learning models. These adversarial attacks can reduce detection accuracy.

8. Real-Time Detection Requirements

Phishing detection must be fast to prevent users from accessing malicious websites. Ensuring high speed and accuracy simultaneously is difficult.

IX. CONCLUSION

The AI-based phishing detection system with Explainable AI (XAI) offers an advanced and reliable approach for identifying and preventing phishing attacks in modern digital environments. The system uses Machine Learning algorithms and Natural Language Processing techniques to analyse URLs, email content, and other relevant features to accurately classify whether an input is phishing or legitimate. Unlike traditional detection methods that rely mainly on blacklists and predefined rules, the proposed system can detect both known and previously unseen phishing attacks by learning patterns from data. This improves detection accuracy and reduces the chances of security breaches.

In addition, the integration of Explainable AI enhances the transparency of the system by providing clear and understandable explanations for each prediction. This helps users and security professionals understand why a particular email or URL is classified as phishing, increasing trust in the system. The system also supports real-time detection, allowing users to take immediate action and avoid potential threats. Overall, the proposed system improves cybersecurity by providing accurate, efficient, and explainable phishing detection, making it a scalable and effective solution for protecting users and organizations from evolving phishing attacks.

X. REFERENCES

- Abdul Basit, Muhammad Zubair, Xiaohong Liu, Athanasios Beznosov, **"Phishing Detection Using Machine Learning Techniques: A Review"**, IEEE, 2019.
- Mohammad A. Alsharnouby, Furkan Alaca, Sezer Sahin, **"A Survey of Phishing Detection Approaches"**, 2015.
- S. Das, A. S. K. Pathan, J. M. Park, **"Deep Learning Techniques for Phishing Detection: A Survey"**, 2020.
- Phish Tank, **"Phishing Data & Resources"**, https://www.phishtank.com

- Kaggle, **"Phishing Websites Dataset"**, https://www.kaggle.com
- Ribeiro, Marco Tulio, et al., **"'Why Should I Trust You?': Explaining the Predictions of Any Classifier"**, ACM SIGKDD, 2016.
- Lundberg, Scott M., Su-In Lee, **"A Unified Approach to Interpreting Model Predictions"**, NeurIPS, 2017.
- Khonji, M., Iraqi, Y., & Jones, A., **"Phishing Detection: A Literature Survey"**, IEEE Communications Surveys & Tutorials, 2013.
- Basnet, R., Sung, A. H., & Liu, Q., **"Detection of Phishing Attacks: A Machine Learning Approach"**, Soft Computing, 2011.
- Jagadeesan, S., & Karthikeyan, P., **"A Survey on Phishing Attack Detection Using AI Techniques"**, International Journal of Computer Applications, 2020.
- Jain, A., & Gupta, B., **"Advanced Techniques for Phishing Detection: Machine Learning and NLP Approaches"**, Journal of Information Security and Applications, 2021.
- OpenAI, **"Explainable AI Techniques for Security Applications"**, https://openai.com/research, 2023.
- Phish Labs, **"The State of Phishing Attacks: Trends & Reports"**, https://www.phishlabs.com/resources, 2022.