Design and Experimental Evaluation of an AI-Based Multi-Source Phishing Detection System Using Explainable AI

PRAPTI GAIKWAD[1]. DR. SWATI JOSHI[2].

1 Department of computer application, PVG's College of Science and Commerce

2 Research Guide, Department of Computer Application, PVG's College of Science and Commerce

Email_Id : gaikwadprapti26@gmail.com

Abstract - Phishing is a major cybersecurity threat in which attackers use fake emails and websites to steal sensitive user information. Many existing detection systems rely on a single source of data and do not clearly explain their decisions, which limits their effectiveness and reduces user trust.

This paper proposes an AI-based multi-source phishing detection system that analyzes both emails and websites. The system extracts features from text content, URL structures, and metadata, and processes them using machine learning and deep learning models. A Random Forest classifier and a Convolutional Neural Network (CNN) were used to identify phishing attacks.

The system was tested using publicly available datasets, including the PhishTank dataset and the Enron Email Dataset. The CNN model achieved 95% accuracy, outperforming the Random Forest model, which achieved 93% accuracy. To improve transparency, an Explainable AI (XAI) module using SHAP and LIME was included to highlight important features influencing the predictions.

The results show that the proposed approach improves detection accuracy, reduces false positives, and provides clear explanations, making it suitable for real-time phishing detection systems.

Keywords - Phishing Detection, Explainable AI, Machine Learning, Deep Learning, Cybersecurity, Multi-Source Detection, Email Security

1. Introduction

Phishing attacks are one of the most common cybersecurity threats today. In these attacks, hackers trick users using fake emails or websites to steal sensitive information such as passwords, banking details, and personal data. With the increase in online services and digital communication, phishing attacks have become more advanced and difficult to detect. Traditional methods like rule-based and signature-based systems are no longer effective because attackers continuously change their techniques to avoid detection.

Artificial Intelligence (AI) has become a powerful solution for phishing detection. Machine Learning (ML) and Deep Learning (DL) models can process large amounts of data, find hidden patterns, and identify phishing attempts more accurately. However, many existing AI-based systems depend on only one type of data (either email or website) and do not clearly explain how decisions are made, which reduces user trust [7].

To solve these issues, this paper presents an AI-based multi-source phishing detection system. The system combines email content, URL structures, and website features to improve detection accuracy. It also uses Explainable AI (XAI) techniques to make the results understandable for users [1], [3]. The system is tested using publicly available datasets, and the results show that combining multiple data sources with explainability improves both accuracy and transparency.

1.1 What is Phishing?

Phishing is a type of cyber attack where attackers try to trick users into sharing sensitive information such as login details, financial data, or personal information. These attacks usually come in the form of fake emails, messages, or websites that look like they are from trusted sources. Because they rely on human behaviour and trust, phishing attacks are still very common and successful.

1.2 Types of Phishing

Phishing attacks can happen in different ways, such as email phishing, website phishing, spear phishing (targeted attacks), and clone phishing. Attackers may also use SMS (SMiShing) or phone calls (Vishing). Even though the methods are different, the main goal is always to steal sensitive information from users.

1.3 Growth and Impact of Phishing Attacks

Phishing attacks have increased rapidly in recent years and are becoming more complex. These attacks can cause serious problems such as financial loss, identity theft, data leaks, and damage to an organization's reputation. As attackers use more advanced techniques, there is a strong need for better and faster detection systems [2].

1.4 Limitations of Existing Detection Systems

Current phishing detection systems have several weaknesses:

- They often rely on only one type of data (email or website)
- They struggle to detect new or unknown (zero-day) attacks
- They do not clearly explain how decisions are made

These problems reduce both the accuracy of detection and user confidence in the system.

Despite these advancements, most existing systems rely on single-source data and lack transparency in their decision-making process, which reduces their effectiveness in detecting modern phishing attacks.

1.5 Problem Statement

Even though many phishing detection methods have been developed, they still face challenges. Most systems cannot effectively handle multi-source data and often lack interpretability. Deep learning models provide good accuracy but act like "black boxes," making their decisions difficult to understand. In addition, many systems produce false positives and fail to detect new types of attacks. Therefore, there is a need for a system that is accurate, uses multiple data sources, and provides clear explanations.

1.6 Research Objectives

The main objectives of this study are:

1. To develop an AI-based phishing detection system using both email and website data
2. To implement and evaluate machine learning and deep learning models
3. To improve accuracy and reduce false positives using multi-source analysis
4. To use Explainable AI techniques to increase transparency and user trust

1.7 Motivation

There is a growing need for smart phishing detection systems that can work in real-time, use multiple data sources, and provide understandable results. AI-based approaches offer a strong solution to these challenges and can help improve overall cybersecurity systems.

1.8 Problem Justification

Despite many existing phishing detection systems, most rely on single-source data and lack transparency. This reduces their effectiveness in detecting modern phishing attacks. Therefore, a multi-source and explainable system is required.

1.9 System Contribution Points

The main contributions of this work are:
• Development of a multi-source phishing detection system using email and website data
• Integration of machine learning and deep learning models for improved accuracy
• Implementation of Explainable AI techniques for model interpretability
• Experimental evaluation using publicly available datasets

2. Literature Survey

Phishing detection techniques have improved over time as cyber attacks have become more advanced. Earlier methods such as rule-based and signature-based systems are no longer very effective because attackers frequently change their strategies to avoid detection. As a result, researchers have started using Artificial Intelligence (AI), including Machine Learning (ML), Deep Learning (DL), and Explainable AI (XAI), to build smarter phishing detection systems [7].

This section reviews recent research (2022–2025) based on data sources, methods used, performance, and limitations. The goal is to identify gaps in existing work and justify the need for a multi-source and explainable phishing detection system.

Earlier studies (2005–2015) mainly used blacklist databases, heuristic rules, and URL analysis. These methods focused on detecting suspicious links and email patterns but were not flexible enough to handle new and evolving attacks.

Many studies use publicly available datasets such as PhishTank, Enron Email Dataset, and Kaggle. for training and evaluation. However, these datasets are often limited to a single data source, which restricts the effectiveness of detection models in real-world scenarios [10], [11], [12].

2.1 Machine Learning Approaches

Machine learning techniques such as Support Vector Machines (SVM), Random Forest, and Decision Trees are commonly used for phishing detection. These models classify emails and websites using features like URL length, domain details, and text patterns. They are fast and require less computational power [2].

However, these methods have some limitations:

- They depend on manual feature selection
- They usually work with only one type of data
- They do not clearly explain their decisions

Due to these limitations, machine learning alone is not sufficient for modern phishing detection systems.

2.2 Deep Learning Approaches

Deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Transformer-based models are widely used for phishing detection. These models can automatically extract important features from email text and website content and often provide higher accuracy than traditional ML methods [6].

Despite their advantages, they have some challenges:

- They behave like “black boxes” and are hard to interpret
- They require high computational resources
- They are often applied to single-source data

These limitations show the need for more transparent and efficient systems.

2.3 Explainable AI (XAI) in Phishing Detection

Explainable AI techniques like LIME and SHAP are used to understand how AI models make decisions. These methods highlight important features that influence predictions, helping users trust the system [3].

However, XAI also has some drawbacks:

- It is often used with only one type of data (email or URL)
- It can increase processing time, affecting real-time performance

Therefore, XAI should be combined with efficient models to be practically useful.

2.4 Multi-Source Detection Systems

Recent studies focus on combining multiple data sources, such as email and website information, to improve phishing detection. These systems often use a mix of ML, DL, and XAI techniques for better performance [1].

However, current multi-source systems still face challenges:

- Lack of datasets that include both email and website data
- Difficulty in integrating different types of data
- Limited support for real-time detection

These issues highlight the need for a more advanced and scalable system.

2.4.1 Recent Studies Relevant to This Work

Several recent studies support the direction of this research:

- Pandya et al. (2025) combined email and website features with XAI techniques to improve transparency [1].
- Sakhare et al. (2025) used advanced ML models for real-time malicious URL detection [2].
- The EXPLICATE framework introduced explainable and multi-source phishing detection using ML and LLMs [3].

- Recent frameworks also explore privacy-aware and multi-agent approaches for phishing detection [4], [5].

These studies show a clear trend toward AI-based, multi-source, and explainable phishing detection systems.

Recent research shows a clear shift toward hybrid systems that combine machine learning, deep learning, and explainable AI for more accurate and trustworthy phishing detection.

2.5 Comparative Analysis of Existing Phishing Detection Studies

Table 1 presents a comparison of different phishing detection methods based on data sources, techniques, real-time capability, explainability, strengths, and limitations.

Table 1: Comparative Analysis of Existing Phishing Detection Studies

Author(s) / Year	Data Source	Method / Technique	Real-Time	Explainability	Strengths	Limitations
Arshad et al., 2023	Emails	BERT (DL)	Moderate	Low	High accuracy for emails	Single-source, lacks transparency
Mohammed et al., 2024	Websites	Random Forest + SVM (ML)	Yes	Low	Fast detection	Single-source, manual feature engineering
Sharma & Gupta, 2022	Emails	CNN + NLP	No	Low	Captures textual patterns	Black-box, no XAI
EXPLICATE, 2025	Emails + Websites	ML + LIME + SHAP + LLM	Moderate	High	Explainable, multi-source	Computationally heavy
Ali et al., 2025	Websites	Autoencoder + Random Forest	No	Moderate	Unsupervised anomaly detection	Limited real-time deployment

The comparison highlights that while deep learning models achieve higher accuracy, they often lack interpretability. On the other hand, machine learning models are faster but depend on manual feature engineering. Multi-source and explainable systems provide better performance but introduce computational complexity.

Despite significant progress in AI-based phishing detection, existing approaches still face several limitations that need to be addressed.

2.6 Identification of Research Gaps

Based on the literature review, the following gaps are identified:

1. Most systems rely on a single data source (email or website)
2. Deep learning models lack interpretability
3. Many studies use limited or outdated datasets
4. Real-time multi-source detection is still challenging

To overcome these issues, this study proposes a multi-source AI-based phishing detection system that combines machine learning, deep learning, and explainable AI. The proposed system aims to improve accuracy, provide clear explanations, and support real-time detection.

This research gap motivates the development of a hybrid system that combines accuracy, efficiency, and interpretability in a unified framework.

3. Methodology

This section explains how the proposed AI-based phishing detection system is designed, implemented, and tested. The system combines email and website data and uses machine learning, deep learning, and Explainable AI (XAI) techniques to detect phishing attacks and provide clear results.

3.1 System Overview

The system follows a step-by-step process that includes data collection, preprocessing, feature extraction, model training, classification, and explanation. It is designed to handle both email and website data together, which improves detection accuracy. This approach also makes the system flexible and suitable for real-world use.

3.2 Data Collection

The system uses publicly available datasets for training and testing:

- PhishTank – provides phishing and legitimate URLs
- Enron Email Dataset – used for email analysis
- Kaggle phishing dataset – used for website-related features

Dataset Summary:

- Total samples: Approximately 10,000 instances collected from multiple datasets

- Training set: 80%
- Testing set: 20%
- Data types: Email text, URL data, and website features

These datasets provide different types of information, which helps in building a multi-source detection system.

3.3 Data Preprocessing

Before training the models, the data is cleaned and prepared:

- Removing duplicate and unnecessary data
- Cleaning text (removing special characters and stop words)
- Normalizing URLs
- Converting text into tokens and numerical form

Feature scaling and normalization are also applied where needed to improve performance.

3.4 Feature Extraction

Different types of features are extracted from the data:

- **Lexical Features:** URL length, special characters, domain patterns
- **Content Features:** Email text analysis using NLP
- **Structural Features:** HTML structure of web pages
- **Metadata Features:** Sender details, domain age, SSL certificate

These features help the models understand patterns related to phishing behavior.

3.5 Model Implementation

Two models are used in the system:

- **Random Forest (Machine Learning)**

  This model works well with structured data and uses multiple decision trees to improve accuracy and reduce overfitting. It is fast and reliable.

- **Convolutional Neural Network (Deep Learning)**

  CNN is used to detect complex patterns in text data. It processes text after converting it into numerical form using embedding techniques.

Training Settings:

Table 2: Training Configuration of Models

Parameter	Value
Train-Test Split	80% – 20%
Batch Size	32
Epochs	10
Optimizer	Adam
Loss Function	Binary Cross-Entropy

The above configuration was used to train and evaluate the machine learning and deep learning models. These parameters were selected to balance performance and computational efficiency.

3.6 Multi-Source Data Integration

To improve performance, the system combines features from email and website data into a single input. This integration allows the model to analyze multiple aspects of phishing attacks at the same time, leading to better accuracy and fewer false positives.

3.7 Explainable AI (XAI) Module

To make the system more transparent, XAI techniques are used:

- SHAP (SHapley Additive Explanations)
- LIME (Local Interpretable Model-Agnostic Explanations)

These methods highlight important features, such as suspicious words, unusual URLs, or metadata patterns, helping users understand why a result is classified as phishing or legitimate.

3.8 Evaluation Metrics

The system is evaluated using standard performance metrics:

- Accuracy
- Precision
- Recall
- F1-score

These metrics help measure how well the system detects phishing attacks, especially in cases where data may be unbalanced.

3.9 System Workflow

The overall working of the system is as follows:

1. Input data (email or website) is collected
2. Data is cleaned and preprocessed
3. Features are extracted
4. Models (Random Forest and CNN) classify the data
5. XAI methods explain the results
6. Final output is generated (Phishing or Legitimate with explanation)

This step-by-step process ensures accurate detection along with clear explanations, making the system suitable for real-world applications.

4. Proposed System

This section explains the design and working of the proposed AI-based phishing detection system. The system analyzes multiple data sources such as emails, URLs, and website content using machine learning (ML), deep learning (DL), and Explainable AI (XAI) techniques.

Unlike traditional systems that use only one type of data, this system combines multiple sources to improve detection accuracy and reduce false results. It also provides clear explanations for its predictions, making it more reliable and user-friendly. The architecture is modular and scalable, which allows it to handle different types of data and adapt to new phishing techniques.

4.1 System Overview

The proposed system adopts a multi-source approach by combining email analysis, website analysis, and URL evaluation to detect phishing attacks more effectively. Instead of relying on a single type of data, the system collects and processes information from multiple sources, which helps in identifying hidden patterns that may not be visible when using only one source. This improves overall detection accuracy and reduces the chances of false predictions.

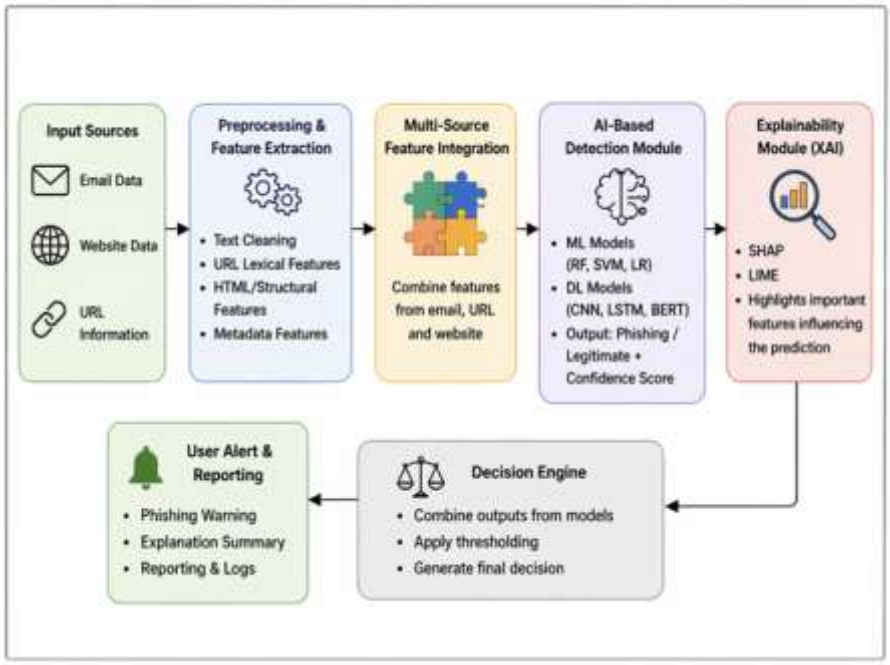

Figure 1: Block/Flow Diagram of the Proposed Framework

The **Figure 1** presents the block diagram of the proposed AI-based multi-source phishing detection framework. It shows the flow of data from input sources through preprocessing, feature extraction, and feature integration stages. The processed data is analyzed using machine learning and deep learning models to classify phishing and legitimate instances. An Explainable AI module is incorporated to interpret model predictions. The final decision is generated by the decision engine and communicated to the user through an alert and reporting module.

The system consists of the following main modules:

1. Input Sources (Email, Website, URL)
2. Preprocessing and Feature Extraction
3. AI-Based Detection (ML and DL models)
4. Explainability Module (SHAP, LIME)
5. Decision Engine
6. User Alert and Reporting

Each module works in coordination with the others to ensure accurate, efficient, and reliable phishing detection. This modular design also makes the system easy to scale and integrate with real-world applications.

4.2 Architecture Description

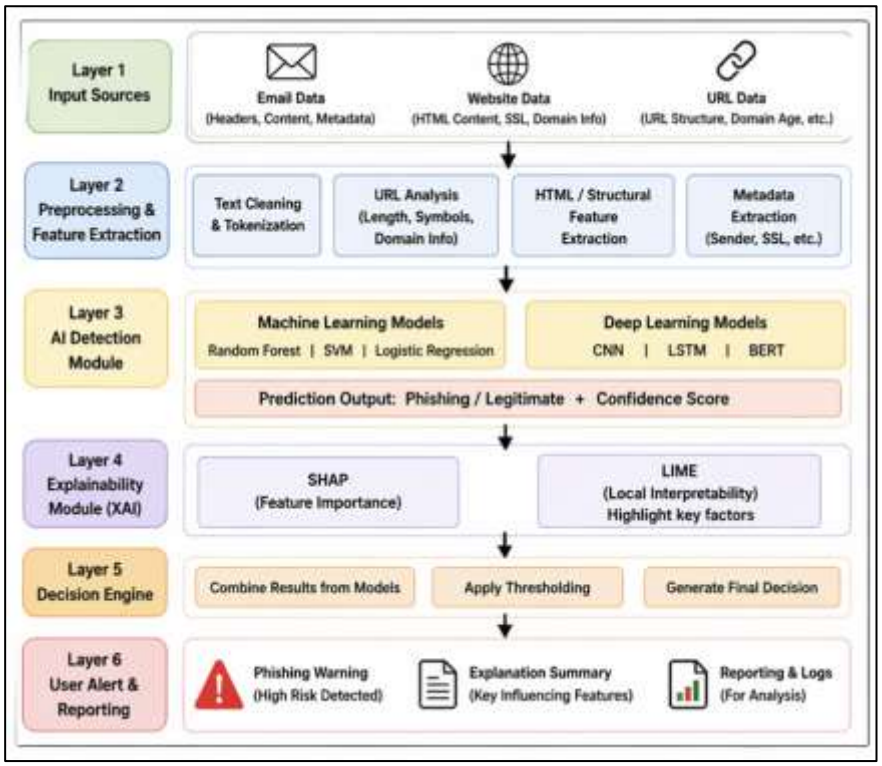

Figure 2: Layered Architecture of the Proposed Phishing Detection System

The figure presents a layered view of the proposed phishing detection system, highlighting the flow from data input and preprocessing to AI-based detection, explainability, decision-making, and user alert generation.

The process begins with collecting input data from different sources such as emails, websites, and URLs. This data is then passed through a preprocessing stage, where it is cleaned and formatted to remove noise and inconsistencies. After preprocessing, important features are extracted, including textual content, URL characteristics, and website structure.

These extracted features are then provided to the AI-based detection module. Machine learning and deep learning models analyze the data and generate a prediction indicating whether the input is phishing or legitimate. Along with the prediction, the system also produces a confidence score, which indicates how certain the model is about its decision.

Once the prediction is made, the Explainable AI module comes into action. It identifies and highlights the key factors that influenced the model's decision, such as suspicious keywords, unusual URL patterns, or abnormal website structures. This helps users understand the reasoning behind the classification.

Finally, the decision engine combines all outputs and generates the final result. The result is then displayed to the user through the alert and reporting module, along with a clear explanation.

System Layers

Layer 1: Input Sources

- Email content, headers, and metadata
- Website URLs, HTML content, SSL details
- Optional user behavior data

This layer collects all necessary data from different sources for analysis.

Layer 2: Preprocessing & Feature Extraction

- Text cleaning and tokenization
- URL analysis (length, symbols, domain age)
- Website structure analysis

This layer prepares the data and extracts meaningful features required for model training and prediction.

Layer 3: AI Detection

- ML models: Random Forest, SVM, Logistic Regression
- DL models: CNN, LSTM, Transformer
- Output: Phishing or Legitimate with confidence score

This is the core layer where intelligent models analyze the data and perform classification.

Layer 4: Explainability

- Uses SHAP and LIME
- Highlights important features affecting the result

This layer improves transparency by explaining how the model arrived at a particular decision.

Layer 5: Decision Engine

- Combines outputs from models
- Generates final prediction

This layer ensures that the final decision is accurate and consistent.

Layer 6: User Alert

- Displays warning messages
- Shows explanation to the user

This layer communicates the results clearly to the end user.

This layered architecture ensures that the system works efficiently and can handle complex phishing detection tasks in a structured manner. It also supports scalability and can be extended for real-time applications.

4.3 Module-Wise Explanation

4.3.1 Email Analysis Module

This module focuses on analyzing email-related data such as the subject line, message content, sender information, and metadata. Natural Language Processing (NLP) techniques are used to understand the text and identify suspicious patterns, such as phishing keywords or unusual writing styles. Deep learning models further enhance detection by capturing complex patterns in the email content.

4.3.2 Website and URL Analysis Module

This module examines URLs and website-related features to detect malicious or fake websites. It analyzes URL structure, domain characteristics, SSL certificate details, redirection behavior, and webpage content. These features help in identifying spoofed or unsafe websites commonly used in phishing attacks.

4.3.3 AI-Based Detection Module

This module is the core component of the system and combines both machine learning and deep learning techniques.

- **Random Forest** is used for structured feature-based classification
- **CNN** is used to extract deeper patterns from textual data

By combining these approaches, the system can capture both simple statistical patterns and complex contextual relationships. The performance of ML and DL models is also compared to determine the most effective approach for phishing detection.

4.3.4 Explainability Module (XAI)

This module uses SHAP and LIME techniques to interpret model predictions. It highlights the most important features that influenced the classification, such as suspicious words in emails or unusual URL structures. This improves transparency and helps users trust the system's decisions.

4.3.5 Decision Engine

The decision engine collects outputs from different models and modules, combines them, and produces the final classification result. It also assigns a confidence score, which helps in understanding the reliability of the prediction.

4.4 System Workflow

The complete workflow of the system is as follows:

1. Input data is collected from email, website, or URL sources
2. The collected data is preprocessed and cleaned
3. Relevant features are extracted from the data
4. Machine learning and deep learning models analyze the features
5. Predictions are explained using XAI techniques
6. Final decision is generated by the system
7. Output is displayed to the user along with explanation

This workflow ensures that the system not only detects phishing attacks accurately but also provides clear and understandable results, making it suitable for real-world cybersecurity applications.

4.5 Pseudocode for Phishing Detection

Algorithm: Multi-Source AI Phishing Detection

Input: Email data (E), URL (U), Website content (W)

Output: Classification Result, Confidence Score, Explanation

1. Receive input (email or URL)
2. Preprocess data (clean text, normalize URLs)
3. Extract features (email, URL, website)
4. Apply ML/DL models for classification
5. Generate prediction score
6. If score > threshold:

   Label = Phishing

   Else:

   Label = Legitimate
7. Apply SHAP/LIME for explanation
8. Combine results in decision engine
9. Generate user alert
10. Display result with explanation

The classification threshold is experimentally set (e.g., 0.5) to balance precision and recall for optimal detection performance.

4.6 Advantages of the Proposed System

- Uses multiple data sources for better accuracy
- Provides explanations using XAI
- Combines ML and DL for better performance
- Scalable and suitable for real-world use
- Can adapt to new phishing techniques

4.7 Future Enhancements

- Use larger combined datasets
- Develop real-time detection tools
- Create browser extensions
- Apply reinforcement learning
- Integrate with email security systems

4.8 Implementation Details

The system was implemented using Python in a cloud environment (Google Colab).

- Scikit-learn → Machine learning models
- TensorFlow/Keras → Deep learning models
- Pandas & NumPy → Data processing
- Matplotlib → Visualization

To handle class imbalance in phishing datasets, techniques such as data balancing or appropriate evaluation metrics (precision, recall, F1-score) are considered to ensure reliable model performance.

The system successfully combines ML, DL, and XAI into one framework. It processes multi-source data efficiently and provides accurate and explainable results, making it useful for real-world cybersecurity applications.

The system is designed to support near real-time detection by processing input data efficiently through optimized preprocessing and lightweight model inference.

The system ensures that sensitive user data is handled securely and can be extended with privacy-preserving techniques in future implementations.

5.Results & Analysis

This section presents the results obtained from testing the proposed AI-based multi-source phishing detection system. The system was evaluated using publicly available datasets such as PhishTank for URL-based detection and Enron Email Dataset for email analysis.

The performance of the models was measured using standard evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics help in understanding how well the system detects phishing attacks.

5.1 Experimental Results

The performance of the machine learning and deep learning models is shown in Table 3.

Table 3: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	93%	92%	91%	91.5%
CNN	95%	94%	93%	93.5%

The results show that the proposed multi-source approach improves detection performance compared to traditional single-source methods.

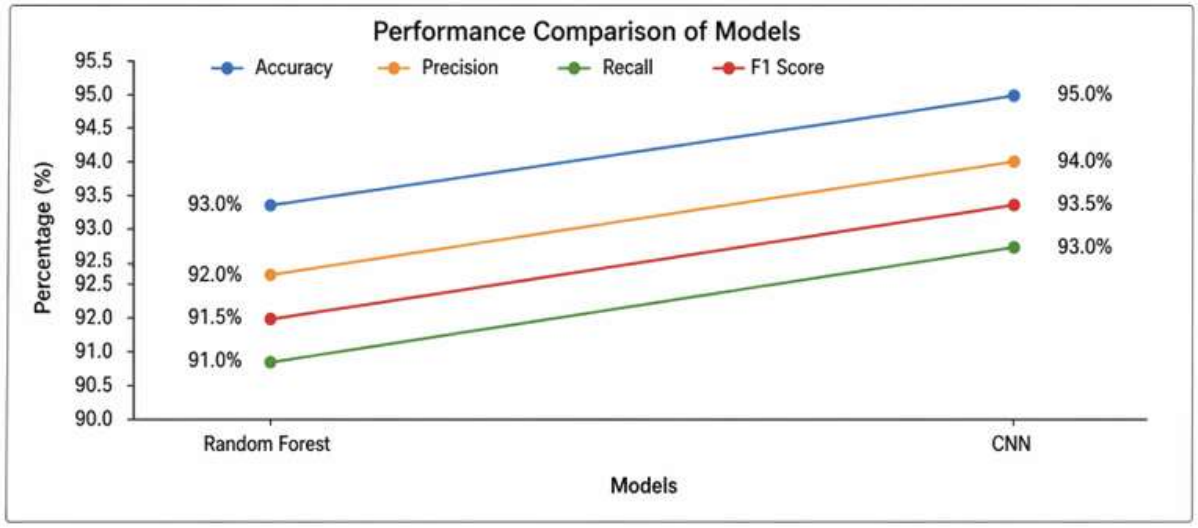

Figure 3: Performance Comparison of Machine Learning and Deep Learning Models

From the results, it is clear that the CNN model performs better than the Random Forest model in terms of accuracy, precision, recall, and F1-score. This is because deep learning models can learn complex patterns from textual data more effectively, especially in email content.

However, the Random Forest model also performs well, particularly when working with structured data such as URLs and metadata. It is more efficient and requires less computational power compared to deep learning models.

Overall, the CNN model shows an improvement of about 2% in accuracy and around 2–3% in other metrics compared to Random Forest. This indicates that deep learning is more effective for detecting complex phishing attacks.

The CNN model performs better because it can automatically learn complex patterns from textual data. It captures contextual relationships in email content more effectively, which improves detection accuracy and reduces misclassification.

5.2 Confusion Matrix Analysis

Table 4: Confusion Matrix for Random Forest Model

	Predicted Phishing	Predicted Legitimate
Actual Phishing	465 (TP)	35 (FN)
Actual Legitimate	35 (FP)	465 (TN)

Table 5: Confusion Matrix for CNN Model

	Predicted Phishing	Predicted Legitimate
Actual Phishing	**475 (TP)**	**25 (FN)**
Actual Legitimate	**25 (FP)**	**475 (TN)**

The confusion matrix is used to evaluate the classification performance of the models by comparing actual and predicted values.

For the Random Forest model, the confusion matrix shows that most phishing and legitimate instances are correctly classified, with a small number of misclassifications.

Similarly, the CNN model demonstrates better performance with fewer false positives and false negatives, indicating improved detection capability.

The comparison shows that the CNN model reduces classification errors compared to the Random Forest model, which supports its higher accuracy and better overall performance.

5.3 Discussion

The experimental results provide several important observations:

- Combining machine learning and deep learning improves detection performance by using both structured and unstructured data
- The CNN model achieves higher accuracy and recall, making it better for detecting complex phishing attacks
- The Random Forest model provides stable performance and is suitable for faster processing
- Using multiple data sources (email, URL, and website features) improves accuracy compared to single-source systems
- Explainable AI techniques (SHAP and LIME) help in understanding model decisions, increasing transparency and user trust
- Compared to existing single-source phishing detection methods, the proposed multi-source approach provides improved accuracy and better interpretability.

5.4 Limitations of Proposed System

Although the proposed system shows strong performance, it has some limitations. The system depends on labelled datasets for training, and deep learning models require higher computational resources. Additionally, real-time deployment on large-scale data remains a challenge.

6. Conclusion

This study presents an AI-based multi-source phishing detection framework that integrates email analysis, website evaluation, and Explainable AI (XAI) techniques to enhance detection accuracy and user trust. Unlike traditional approaches that rely on a single data source, the proposed system utilizes multiple inputs such as email content, URLs, and website features, enabling more effective identification of both simple and complex phishing attacks. The integration of machine learning and deep learning models allows the system to learn patterns from data and continuously improve its performance.

A key contribution of this work is the incorporation of Explainable AI methods, including SHAP and LIME, which provide insights into the factors influencing model decisions. This improves transparency and helps users understand the reasoning behind each classification. Experimental results demonstrate that the deep learning model (CNN) effectively captures complex textual patterns, while the Random Forest model offers stable performance for structured data. Together, they form a balanced and reliable detection system.

Despite these advantages, the system depends on high-quality labeled data and requires significant computational resources for deep learning models. Real-time deployment at scale also remains a challenge. Overall, the proposed framework enhances both detection performance and interpretability, making it suitable for practical cybersecurity applications.

Future Work

Although the proposed system shows strong performance, there are several areas where further improvements can be made.

- The system can be extended to support real-time phishing detection, allowing it to analyze live data and provide instant alerts to users.
- Larger and more diverse datasets can be used to train the models, which will improve their ability to detect new and evolving phishing attacks.
- The system can be integrated with email clients and web browsers to provide continuous protection against phishing threats in real-world environments.
- Advanced deep learning models, such as transformer-based architectures, can be explored to further improve detection accuracy and handle complex language patterns.
- Reinforcement learning techniques can be applied to enable the system to adapt dynamically to new attack strategies and continuously improve its performance over time.

These future enhancements will help in developing a more robust, scalable, and intelligent phishing detection system that can effectively handle modern cybersecurity challenges.

7. References

[1] P. Pandya *et al.*, “Combined email and website features with XAI techniques for improved phishing detection,” unpublished preprint, 2025.

[2] S. Sakhare *et al.*, “Real-time malicious URL detection using advanced ML models,” unpublished preprint, 2025.

[3] B. Lim, R. Huerta, A. Sotelo, A. Quintela, and P. Kumar, “EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability,” *arXiv preprint* arXiv:2503.20796, 2025. [Online]. Available: https://arxiv.org/abs/2503.20796

[4] W. Li, S. Manickam, Y.-W. Chong, and S. Karuppayah, “PhishDebate: An LLM-based multi-agent framework for phishing website detection,” *arXiv preprint* arXiv:2506.15656, 2025. [Online]. Available: https://arxiv.org/abs/2506.15656

[5] W. Kang, N. Wang, J. Seung, S. Wang, and A. Abuadbba, "EPhishCADE: A privacy-aware multidimensional framework for email phishing campaign detection," *arXiv preprint* arXiv:2502.20621, 2025. [Online]. Available: https://arxiv.org/abs/2502.20621

[6] M. Murhej and G. Nallasivan, "Multimodal framework for phishing attack detection and mitigation through behavior analysis using EM-BERT and SPCA-based EAI-SC-LSTM," *Frontiers in Communications and Networks*, 2025. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frcmn.2025.1587654/full

[7] D. Popescul and L. D. Radu, "AI in phishing detection: A bibliometric review," *Frontiers in Artificial Intelligence*, 2025. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1496580/full

[8] M. Kshirsagar, V. Rathi, and C. Ryan, "Meta-learner based frameworks for interpretable email spam detection," *Frontiers in Artificial Intelligence*, 2025. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1569804/full

[9] A. Al Helali, W. Maqableh, H. Fakhouri, and W. Alkhadour, "FedPhishLLM: A privacy-preserving and explainable phishing detection mechanism using federated learning and LLMs," *Journal of King Saud University – Computer and Information Sciences*, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s44443-025-00267-0

[10] PhishTank, "PhishTank: An open phishing database," 2024. [Online]. Available: https://www.phishtank.com

[11] W. W. Cohen, "Enron email dataset," Carnegie Mellon University, 2015. [Online]. Available: https://www.cs.cmu.edu/~enron/

[12] Kaggle, "Phishing websites dataset," 2023. [Online]. Available: https://www.kaggle.com