

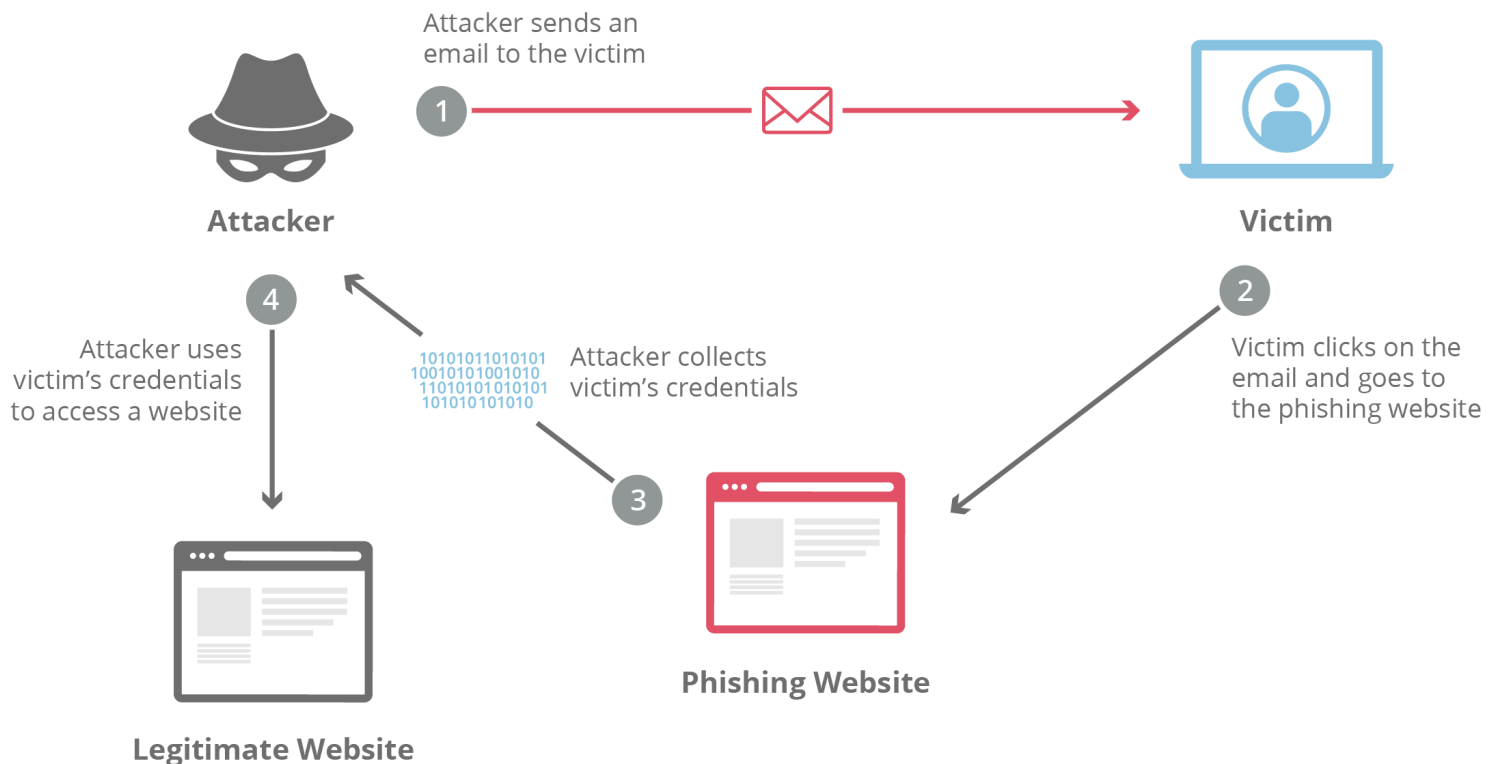
AI to detect Phishing

Made By Dev Mandora , Harshal Mehta , Krisha Mehta

Introduction

The internet has made it easier than ever to connect and do business. But it's also given bad guys more ways to trick people. Phishing is a common online scam where criminals try to get your personal information, like passwords and bank details. They often pretend to be someone you trust – maybe your bank, a popular online store, or even a friend. They might send you an email that looks official, with logos and convincing language, asking you to click a link or open an attachment. This link often takes you to a fake website that looks just like the real thing, where they try to steal your login information. These attacks can cause a lot of damage, from stealing money directly from your bank account to leaking private data that can be used for identity theft. Phishing is a serious problem because it affects millions of people and businesses every year.

Let's take an example of how someone might encounter a phishing attack. The following image illustrates the step-by-step process of how phishing works in this scenario



How a Phishing Attack Works

While this illustrates a common phishing technique, the landscape of these attacks is vast and ever-changing. From deceptive emails to malicious websites and SMS messages, how can we possibly keep up? Artificial intelligence offers a powerful and adaptive approach, with the potential to not only detect current threats but also anticipate and neutralize new and evolving phishing tactics

How AI Outperforms Traditional Detection Methods


Traditional phishing detection methods, such as blacklists and rule-based systems, have long been used to combat these threats. However, they struggle to detect emerging attacks that bypass predefined security measures. AI-powered systems provide an adaptive and proactive approach to phishing detection. The following table compares these methods:

Method	Advantages	Disadvantages
Blacklist-based	Simple, fast detection of known phishing sites	Fails to detect new, unknown phishing domains
Rule-based	Effective for previously identified threats	Requires continuous manual updates, vulnerable to novel phishing techniques
Signature-based	Quick identification of threats with known patterns	Cannot detect zero-day phishing attacks or sophisticated social engineering scams
AI-Based	Adaptive learning, real-time detection, and behavioral analysis	High resource and computational requirements

Unlike static detection methods, AI continuously learns from new threats, adapting in real time. This makes AI an essential tool in combating phishing attacks that are constantly evolving.

Case Study : The 2020 Twitter Bitcoin Scam

One of the most high-profile phishing incidents occurred in July 2020 , when cybercriminals successfully targeted Twitter employees through a spear-phishing attack . The attackers used social engineering techniques to gain access to Twitter’s internal systems, enabling them to hijack verified accounts of prominent figures, including Barack Obama, Elon Musk, and Apple. The compromised accounts tweeted fraudulent messages, promising to double any Bitcoin sent to a specified wallet. Within hours, the scam collected over \$118,000 in Bitcoin from unsuspecting victims.



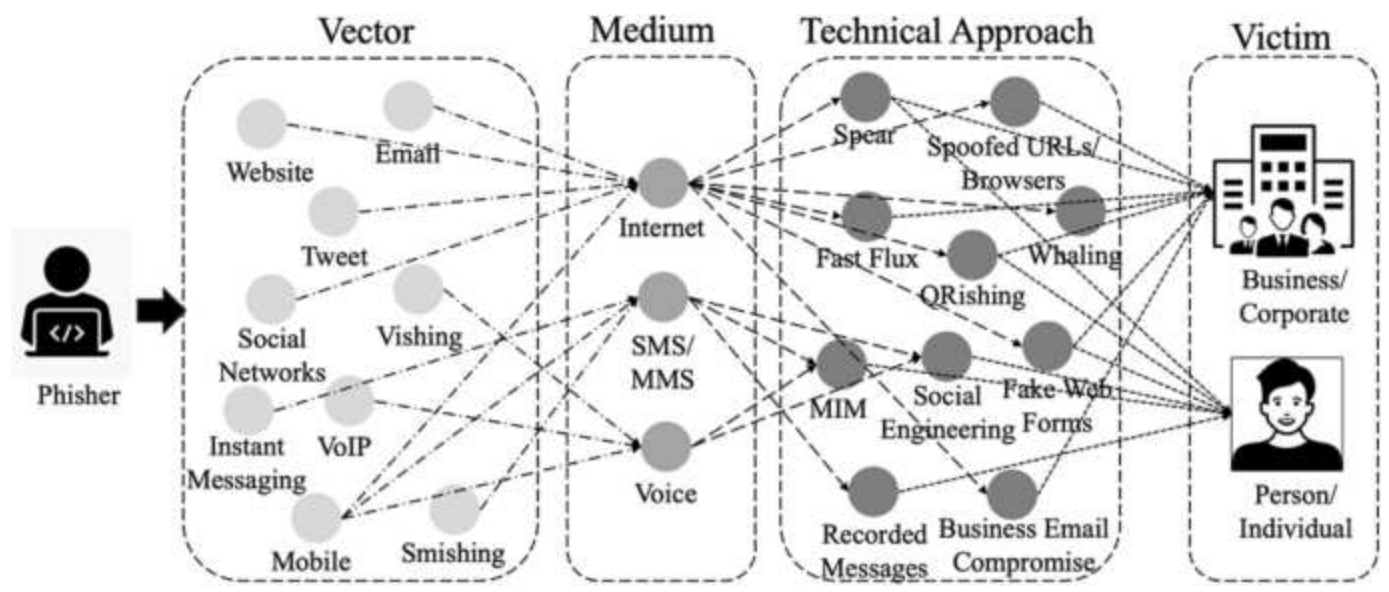
This incident highlights the growing sophistication of phishing attacks, where cybercriminals exploit human vulnerabilities rather than technical flaws. It also underscores the importance of AI-driven phishing detection in monitoring internal communications and identifying anomalies in real-time.

1. Current Progress in AI for Detecting Phishing

Machine Learning (ML) and Deep Learning (DL) are enhancing AI systems to detect phishing attacks by processing large amounts of data, identifying patterns, and learning behaviors that signify phishing attempts. These technologies detect patterns in content, such as emails, text messages, social media posts,

and multimedia files. Similarly, Natural Language Processing (NLP) techniques enable AI to understand the language used in phishing attempts, helping to identify social engineering tactics. This is especially crucial when attackers mimic the writing style of legitimate senders to deceive users.

Additionally, AI has expanded beyond detecting phishing solely in emails. It now analyzes multiple communication channels, including social media posts, SMS messages, and mobile apps, providing a more comprehensive defense against phishing threats.



Phishing Attack Vectors and Techniques

This figure illustrates the diverse attack vectors employed in phishing, categorized by medium (e.g., website, social networks), technical approach (e.g., spear phishing, fast flux), and victim type (business/corporate or person/individual). It highlights the multi-channel nature of phishing attacks, moving beyond email to include various internet-based and mobile communication methods, each utilizing distinct techniques to target different victims. The dashed lines suggest the potential intersectionality of these attack vectors and methods.

To better understand how AI detects phishing, the table below summarizes various detection techniques

Methods	Details	Key Points	Advantages	Disadvantages
Phishing Attacks: Overview and Evolution	Definition and Types of Phishing Attacks	Various types include email phishing, spear phishing, whaling, vishing, and smishing.	Understanding attack types helps tailor defense strategies.	Attackers constantly evolve tactics, making it challenging to keep defenses up to date.

	Evolution of Phishing Techniques	Phishing has evolved from simple scams to sophisticated, multi-channel attacks using advanced tactics.	Recognizes the need for adaptive and robust security measures.	Increased sophistication makes detection harder and more resource-intensive.
	Impact on Individuals and Organizations	Phishing can lead to identity theft, financial loss, data breaches, and reputational damage.	Highlights the critical importance of effective phishing prevention to protect users and organizations.	Significant financial and reputational impact can occur if defenses fail.
Traditional Phishing Detection Methods	Signature-Based Detection	Compares incoming data with known phishing signatures but struggles with new, unknown threats.	Effective against known threats with established signatures.	Ineffective against zero-day attacks and novel phishing techniques.
	Heuristic-Based Detection	Uses rules to identify suspicious patterns but can result in high false positives.	Can detect novel threats by analyzing suspicious behavior patterns.	High false-positive rate can overwhelm security teams with unnecessary alerts.
	Limitations of Traditional Approaches	Traditional methods are reactive, often rigid, and struggle with novel phishing tactics.	Provides a foundation for developing more advanced, adaptive security measures.	Limited adaptability and slow to respond to evolving threats.
AI-Based Phishing Detection	Overview of AI and ML in Cybersecurity	AI/ML can analyze large datasets, identify complex patterns, and adapt to new threats.	Offers real-time, scalable detection with the ability to learn and adapt to new threats.	Requires significant computational resources and large datasets for effective training.

	Supervised, Unsupervised, and Reinforcement Learning	Different learning approaches are used, with each offering unique benefits for phishing detection.	Provides flexibility in detection methods, catering to various types of data and threat scenarios.	Supervised learning depends on high-quality labeled data; unsupervised learning can be less accurate.
	Comparative Analysis of AI Models	AI models like logistic regression, CNNs, and RNNs outperform traditional methods in phishing detection.	Higher accuracy, adaptability, and efficiency in detecting complex and evolving phishing threats.	Complex models can be opaque, making it difficult to interpret results and understand decision-making.
Multi-Channel Communication Platforms	Overview of Communication Channels	Channels include email, SMS, social media, each with unique phishing risks.	Ensures comprehensive security across diverse communication platforms.	Managing security across multiple channels can be complex and resource-intensive.

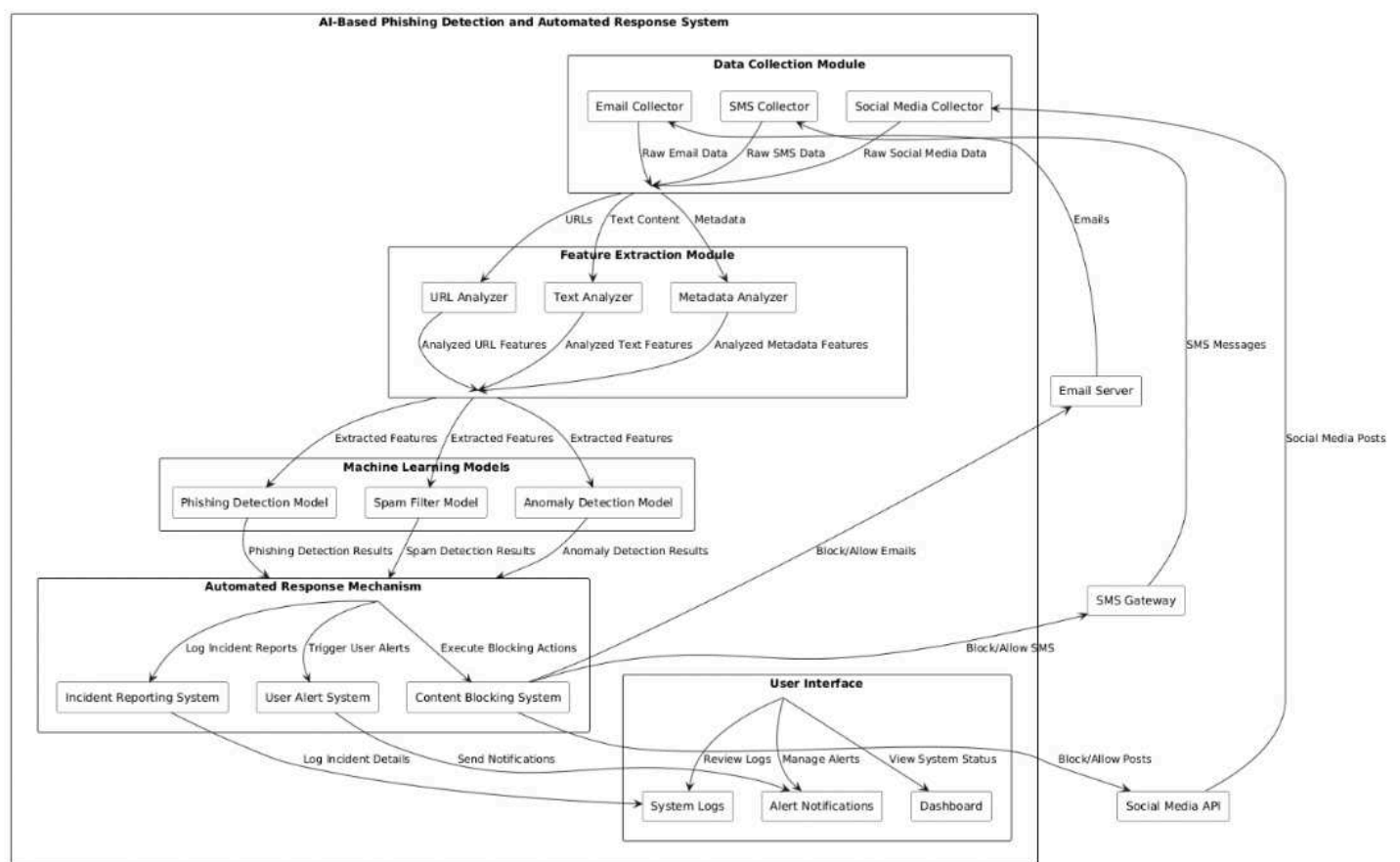
AI-driven automated response systems offer several key advantages in combating phishing attacks. Real-time mitigation capabilities allow these systems to trigger immediate responses upon detecting suspicious activity, such as quarantining or deleting malicious messages before they reach potential victims. Simultaneously, automated user notifications alert individuals to potential phishing threats, ensuring timely awareness. Furthermore, these systems can automatically initiate predefined incident response protocols, which include logging the attack, generating reports, and blocking the phishing source. The overall impact on cybersecurity is a decrease in successful phishing attacks, as organizations benefit from the adaptive nature of AI-based defenses that learn and respond to evolving phishing techniques.

2. Current Architecture in AI-Based Phishing Detection Systems

AI-driven phishing detection systems employ a multi-layered approach to identify and mitigate phishing threats across multiple channels. Initially, these systems collect data from a variety of communication platforms, including email, SMS messages, social media posts, and mobile applications, ensuring comprehensive coverage and a holistic view of potential attack vectors. Following data collection, the system performs feature extraction, identifying key indicators of phishing attempts. This process involves analyzing elements such as URL structure (e.g., suspicious domain names, URL shortening), text content (e.g., use of urgency, grammatical errors, emotional appeals), and metadata (e.g., sender information, message headers) to detect subtle signs of malicious intent. The extracted features are then fed into machine learning models, which are trained on extensive datasets of both phishing and legitimate

messages. These models leverage pattern recognition algorithms to classify incoming messages as either phishing or legitimate, continuously learning and adapting to new and evolving phishing techniques.

Upon detection of a phishing attempt, the automated response mechanism triggers predefined actions in real-time. These actions may include blocking or quarantining suspicious content, alerting security personnel, and even initiating user education campaigns. Finally, a user interface (UI) provides security teams with a centralized platform to monitor system activity, review logs of detected phishing threats and actions taken, manage alerts, and fine-tune system parameters, allowing for proactive threat management and continuous improvement of the detection process. Furthermore, advanced systems may incorporate techniques like Natural Language Processing (NLP) to understand the context and sentiment of messages, enhancing the accuracy of phishing detection, particularly in cases involving sophisticated social engineering tactics. Some systems also utilize anomaly detection to identify deviations from normal communication patterns, which can be indicative of a compromised account or a novel phishing attack.



The above diagram illustrates the architecture of the AI-based phishing detection and automated response system, outlining the flow of data from collection across multiple channels (email, SMS, social media) through feature extraction, machine learning-based detection, and automated responses, including a user interface for monitoring and management

Challenges & Limitations of AI-Based Phishing Detection

Despite the significant advancements in AI-driven phishing detection, several challenges and limitations hinder its effectiveness. One major challenge is adversarial phishing attacks, where attackers manipulate email content, URLs, or images to evade detection. For instance, adversarial text modifications involve

slight alterations in phishing messages (e.g., replacing characters with visually similar symbols) to bypass AI filters. Similarly, image-based phishing techniques embed phishing content within images, making it difficult for AI models reliant on textual analysis to detect malicious intent. Another limitation is the issue of false positives and false negatives. AI-based systems may incorrectly classify legitimate emails as phishing (false positives), causing disruption in business communication, or fail to detect well-crafted phishing attempts (false negatives), leading to security breaches. Furthermore, the resource-intensive nature of deep learning-based phishing detection poses scalability concerns. These models require large-scale labeled datasets, continuous training, and significant computational power, making deployment challenging for organizations with limited resources. Additionally, data privacy and ethical concerns emerge due to the reliance of AI models on user data for training. The risk of exposing sensitive information, potential biases in training datasets, and the black-box nature of some AI models raise concerns regarding transparency and trust in phishing detection systems. Addressing these challenges requires a combination of advanced AI techniques, improved dataset curation, and integration with human-in-the-loop verification systems to enhance the robustness of phishing detection.

3. Future Development Possibilities

Future developments in AI-driven phishing detection will focus on improving real-time accuracy, adapting to evolving attack methods, and enhancing usability. One significant area of advancement is the integration of hybrid detection models that combine machine learning with rule-based techniques to improve detection accuracy. Ensemble models, such as a combination of Random Forest and deep learning, can help identify sophisticated phishing attempts by analyzing email structure, writing style, and user interaction patterns. Context-aware detection mechanisms will further enhance the system's ability to distinguish between legitimate and phishing content.

Another critical development is adversarial training, which strengthens AI models against evolving phishing techniques. As attackers continuously refine their methods to evade detection, incorporating adversarial learning strategies can make phishing detection models more resilient. The use of Generative Adversarial Networks (GANs) can simulate new phishing strategies, allowing models to adapt preemptively. Additionally, zero-shot and few-shot learning approaches can help AI systems detect previously unseen phishing techniques with minimal training data.

Real-time phishing detection and automation will also play a crucial role in future systems. The integration of browser extensions and email security plugins will enable immediate phishing alerts, while advancements in natural language processing will enhance deep semantic analysis of phishing content. Automated mitigation mechanisms, such as blocking malicious links in real time, will further minimize potential threats.

Expanding phishing detection capabilities to regional and multilingual content is another important area of development. Current phishing detection models are primarily trained on English datasets, limiting their effectiveness in non-English-speaking regions. Future advancements will involve training AI models on diverse datasets that include phishing emails and websites from multiple languages and dialects, ensuring a more comprehensive defense against global phishing threats.

Finally, improving user awareness and integrating explainable AI (XAI) will make phishing detection systems more transparent and user-friendly. AI-driven interactive assistants can educate users on phishing threats and provide real-time guidance. Explainable AI techniques will allow users to understand why an

email or website is flagged as phishing, increasing trust and adoption. Additionally, user-friendly reporting systems will encourage community-driven phishing detection, allowing users to contribute to the system’s ongoing improvement.

By focusing on these advancements, AI-driven phishing detection systems will become more accurate, adaptive, and accessible, ensuring robust protection against increasingly sophisticated cyber threats.

4. Analysis of Research Papers

With the rise of sophisticated phishing attacks, AI-based detection methods have gained significant attention. Researchers have explored machine learning and deep learning techniques to analyze phishing emails and websites, achieving high accuracy. Models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest have shown promising results in detecting phishing patterns.

Despite their effectiveness, challenges remain, including evolving phishing tactics and the need for real-time detection with minimal false positives. To improve accuracy, studies suggest hybrid models and advanced feature extraction techniques. The following table compares key research papers, highlighting their methodologies, findings, and challenges.

Comparative Analysis of Research Papers on Phishing Detection & News Classification

Parameter	Research Paper 1: "Phishing Attacks Detection: A Machine Learning-Based Approach" (Fatima Salahdine et al.)	Research Paper 2: "Fake News Detection Using Machine Learning" (Rajesh Kumar et al.)	Research Paper 3: "Real-Time News Categorization Using NLP and Deep Learning" (Ananya Verma & Rohit Mishra)	Research Paper 4: "Analysis and Prevention of AI-Based Phishing Email Attacks"	Research Paper 5: "Identification of Phishing Attacks using Machine Learning Algorithm"
Focus Area	Machine learning-based phishing email detection.	Detecting fake news using ML techniques.	Real-time news classification using NLP and DL.	AI-generated phishing emails and their detection.	Phishing website detection using machine learning.

Phishing Methods Discussed	Email-based phishing, social engineering.	N/A (focuses on fake news detection).	N/A (focuses on real-time news classification).	AI-generated phishing emails, email-based phishing.	Emails, fake websites, HTTP phishing, clone phishing, spear phishing, whaling.
Machine Learning Techniques	Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Networks (ANN).	Random Forest, Naïve Bayes, Gradient Boosting, Bi-LSTM.	BERT, LSTM, CNN, TF-IDF vectorization, Attention Mechanism.	MALLET (Naïve Bayes, Max Entropy, Winnow), UDAT (Style-based analysis), LSTM, Ensemble.	Random Forest, XGBoost, Support Vector Machines (SVM).
Dataset Used	4000 phishing emails (University of North Dakota email system).	LIAR dataset (12,800 labeled statements from Politifact).	120,000 news articles from Reuters, BBC, and CNN.	865 AI-generated phishing emails compared with Enron, Nigerian scam, and Ling-Spam datasets.	5000 phishing and 5000 legitimate URLs from PhishTank and other sources.
Evaluation Metrics	Detection Probability (Pd), False Alarm Probability (Pfa), Miss-Detection Probability (Pmd), Accuracy.	Accuracy, Precision, Recall, F1-score, ROC-AUC.	Accuracy, Precision, Recall, BLEU Score.	Accuracy, Precision, Recall, F1-score, Confusion Matrix.	Accuracy, Performance Comparison.
Best Performing Model	ANN with 2 hidden layers (100 neurons each, ReLU)	Bi-LSTM with TF-IDF achieved 95.2% accuracy.	BERT with Attention Mechanism (96.3% accuracy).	MALLET with Naïve Bayes (99.3%	XGBoost (94.2% accuracy).

	achieved 94.5% accuracy.			accuracy), Ensemble (99.5%).	
Key Findings	Phishing emails can be detected with high accuracy using ML classifiers.	Fake news can be detected with high precision using deep learning models.	NLP-based deep learning models enhance real-time news classification.	AI-generated phishing emails differ in writing style, word choice, sentiment, and complexity.	Phishing websites can be detected with high accuracy using ML-based classifiers.
Challenges Discussed	High variability in phishing techniques, balancing accuracy with false positives.	Difficulty in detecting subtle misinformation, evolving fake news patterns.	Imbalanced data, ensuring real-time news filtering.	AI-generated phishing emails constantly evolve, requiring adaptive detection.	Phishers use new evasion techniques, making real-time detection difficult.
Proposed Improvements	Deep learning (CNN, RNN), better feature extraction, larger datasets.	Improving model generalization, fact-checking integration, adversarial training.	Expanding to regional languages, better handling of imbalanced datasets.	Training ML models with AI-generated phishing emails for better detection.	Developing browser extensions or GUI tools for phishing detection.
Future Work	Real-time phishing detection, hybrid models combining ML with rule-based techniques.	Enhancing multilingual fake news detection, better adversarial robustness.	Exploring deep learning models, integrating real-time detection.	AI-generated phishing emails are different from manually created scams, and ML can	ML-based classifiers can detect phishing websites efficiently, but ongoing research is needed for real-time solutions.

detect them
effectively.

Conclusion	ML-based phishing detection is fast and accurate but needs real-time optimization.	Fake news detection improves significantly with deep learning and fact-checking.	AI-based news classification is efficient, enhancing personalized recommendations.	AI-generated phishing emails pose new challenges, but ML models can detect them effectively.	Phishing websites can be detected efficiently with ML, but real-time detection remains a challenge.
-------------------	--	--	--	--	---

References

- Albarqi, A., Alzaid, E., Ghamdi, F., Asiri, S., & Kar, J. (2015). Public key infrastructure: A survey. *Journal of Information Security*, 6(1), 31–37.
- Arjoune, Y., Salahdine, F., Islam, Md., Ghribi, E., & Kaabouch, N. (2020). A novel jamming attacks detection approach based on machine learning for wireless communication. *International Conference on Information Networking*, 1–6.
- Bircano, C., & Arica, N. (2018). A comparison of activation functions in artificial neural networks. *Signal Processing and Communications Applications Conference*, 1–4.
- Chanti, S., & Chithralekha, T. (2020). Classification of anti-phishing solutions. *SN Computer Science*, 1, 1–18.
- Chintale, P., Khanna, A., Korada, L., Desaboyina, G., & Nerella, H. AI-Enhanced Cybersecurity Measures for Protecting Financial Assets.
- Ebubekir, B., Diri, B., & Sahingoz, O.K. (2017). NLP based phishing attack detection from URLs. *International Conference on Intelligent Systems Design and Applications*, Springer, Cham, 608–618.
- Eric, M., Kirda, E., & Kruegel, C. (2008). Visual-similarity-based phishing detection. *Proceedings of the 4th International Conference on Security and Privacy in Communication Networks*, 1–6.
- Gowtham, R., Krishnamurthi, I., & Kumar, K. S. S. (2014). An efficacious method for detecting phishing webpages through target domain identification. *Decision Support Systems*, 61, 12–22.
- Gupta, B., Arachchilage, N., & Psannis, K. (2018). Defending against phishing attacks: Taxonomy of methods, current issues and future directions. *Telecommunication Systems*, 67, 247–267.
- He, J., & Zhu, Y. (2014). Social engineering/phishing. *Encyclopedia of Social Network Analysis and Mining*, 1777–1783.
- Hong, J., Kim, T., & Kim, S. (2020). Phishing URL detection with lexical features and blacklisted domains. *Adaptive and Autonomous Security Cyber Systems*, 253–267.
- Huang, Y., Yang, Q., Qin, J., & Wen, W. (2019). Phishing URL Detection via CNN and Attention-Based Hierarchical RNN. *IEEE International Conference on Trust, Security, Privacy in Computing and Communications*, 112–119.

- Josna, P., Fathima, K.A.F., Gayathri, S., Elias, G.E., & Menon, A.A. (2022). A comparative study of machine learning models for the detection of Phishing Websites. *International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–7.
- Kharraz, A., Robertson, W., & Kirda, E. (2018). Surveyance: Automatically detecting online survey scams. *IEEE Symposium on Security and Privacy*, 723–739.
- Kommisetty, P. D. N. K., & Abhireddy, N. (2024). Cloud Migration Strategies: Ensuring Seamless Integration and Scalability in Dynamic Business Environments. *International Journal of Engineering and Computer Science*, 13(04), 26146–26156.
- Koray, S.O., Buber, E., Demir, O., & Diri, B. (2019). *Expert Systems with Applications**, 117, 345–357.
- Krishnamurthy, S., & Ve, A. (2017). Information retrieval models: Trends and techniques. *Web Semantics: Science, Services and Applications*, 42, 17–42.
- Mahida, A. Secure Data Outsourcing Techniques for Cloud Storage.
- Mahida, A., Chintale, P., & Deshmukh, H. (2024). Enhancing Fraud Detection in Real Time using DataOps on Elastic Platforms.
- Moghimi, M., & Varjani, A. (2016). New rule-based phishing detection method. *Expert Systems with Applications*, 53, 231–242.
- Pillai, S. E. V. S., Avacharmal, R., Reddy, R. A., Pareek, P. K., & Zanke, P. (2024, April). Transductive–Long Short-Term Memory Network for the Fake News Detection. *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, 1–4. IEEE.
- Rachna, D., & Tygar, J.D. (2005). The battle against phishing: Dynamic security skins. *Proceedings of the 2005 Symposium on Usable Privacy and Security*, 77–88.
- Ramesh, G., Krishnamurthi, I., & Kumar, K. (2014). *Decision Support Systems**, 61, 12–22.
- Roy, A., Basak, K., Ekbal, A., & Bhattacharyya, P. (2018). A Deep Ensemble Framework for Fake News Detection and Classification. *arXiv:1811.04670*. <https://arxiv.org/abs/1811.04670>
- Salahdine, F., & Kaabouch, N. (2019). Social Engineering Attacks: A Survey. *Future Internet*, 11(4), 89.
- Ali, M., Amin, A., & Irfan, M. (2023). AI-Driven Phishing Detection: Current Advancements and Future Prospects. *PMJ Journal*. <https://internationalpubls.com/index.php/pmj/article/view/2312>
- Idrees, A., Rajarajan, R., & Conti, S. (2022). AI-Driven Phishing Detection: A Review of Recent Advances. *IEEE Xplore*. <https://ieeexplore.ieee.org/abstract/document/9716113>
- Author(s) Unknown. (n.d.). AI-Driven Phishing Detection Systems. *ResearchGate*. https://www.researchgate.net/publication/382917933_AI-Driven_Phishing_Detection_Systems
- Author(s) Unknown. (n.d.). A Comprehensive Survey of AI-Enabled Phishing Attacks Detection Techniques. *ResearchGate*. https://www.researchgate.net/publication/344583507_A_comprehensive_survey_of_AI-enabled_phishing_attacks_detection_techniques
- Author(s) Unknown. (n.d.). AI-Powered Phishing Detection: Analyzing Trends and Challenges. *DocWorkspace*. <https://in.docworkspace.com/d/sIDrF-qChAqKQib0G?sa=601.1094&ps=1&fn=researchpaper.pdf>