Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Detection of Phishing Websites and Emails Using Explainable Machine Learning Models

Pandya Himani[1], Dr. Khyati Zalawadia[2], Dr. Harish Prajapati[3]
[1]Faculty of Engineering & Technology Parul University, Vadodara, Gujarat, India.
[2]Faculty of Engineering & Technology Parul University, Vadodara, Gujarat, India.
[3] Faculty of Engineering & Technology Parul University, Vadodara, Gujarat, India.
E-Mail:- 2303032010023@ paruluniversity.ac.in, Khyati.Zalawadia29490@paruluniversity.ac.in, harish.prajapati35068@paruluniversity.ac.in

ARTICLE INFO

Received: 24 Dec 2024

Revised: 16 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Phishing attacks pose serious cybersecurity concerns by using phony emails and websites to obtain private data. Intelligent and explicable solutions are required because traditional detection systems are unable to keep up with the latest phishing techniques. This study integrates behavioural analysis, optical character recognition (OCR), and natural language processing (NLP) to develop a phishing detection system based on Explainable Machine Learning (XML) models. The system uses supervised models, including ensemble approaches, for precise classification and anomaly detection for adaptive learning. Model transparency is improved by explainability techniques like SHAP and LIME, which help cybersecurity experts understand decisions. High detection accuracy, adaptability, and increased reliability over traditional methods are demonstrated by the experimental results. The suggested system provides a strong solution for cybersecurity resilience by improving phishing defence through real-time detection, alert systems, and adaptive learning.

Keywords: Phishing detection, Explainable AI, Machine Learning, Natural Language Processing, Optical Character Recognition, Anomaly detection, Supervised learning, Ensemble methods, SHAP, LIME, Cybersecurity.

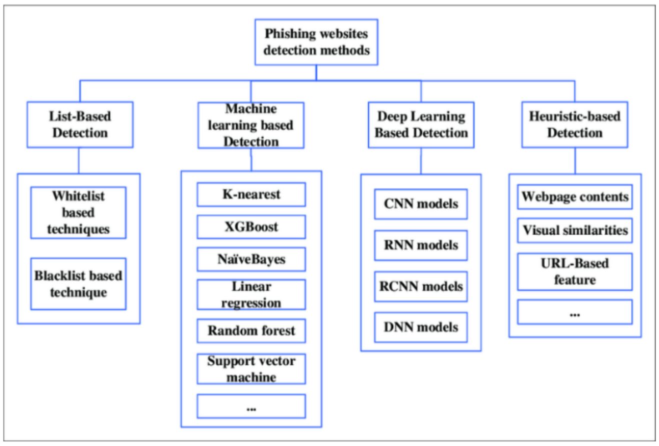
1. INTRODUCTION

One of the most common cyberthreats nowadays is phishing, which takes advantage of technical flaws and human psychology to obtain private information. Cybercriminals pose as trustworthy organizations and trick people into disclosing private information by using phony emails, harmful URLs, and fake websites. Because they can't keep up with changing attack tactics, traditional rule-based detection methods like blacklisting and signature-based approaches frequently miss new phishing attempts.

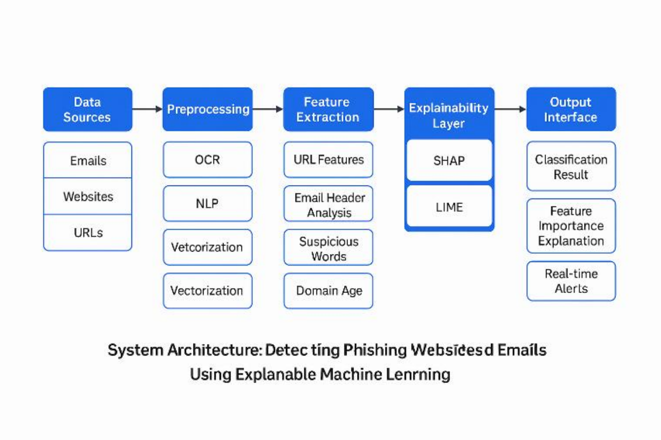
A viable substitute is machine learning (ML), which offers flexible and dynamic ways to identify phishing attempts. But because the majority of ML-based detection models function as "black boxes," it might be difficult to understand how they make decisions. By improving accountability and transparency, explainable AI (XAI) overcomes this constraint and enables cybersecurity experts to comprehend and have faith in model predictions.

This study uses Explainable Machine Learning Models to suggest a new method for identifying phishing emails and websites. To extract important information from input data sources, the system combines sophisticated behavioural analysis, optical character recognition (OCR), and natural language processing (NLP). Both supervised and unsupervised learning strategies are used in the suggested model to increase detection precision and flexibility. Furthermore, explainability strategies like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) offer insights into model predictions, while ensemble learning approaches improve robustness.

This paper explores the proposed methodology, system architecture, and evaluation results, demonstrating the effectiveness of explainable ML models in detecting and mitigating phishing attacks.



(fig.1.1 work flow of phising detection)



(fig.1.2 System Design for Detecting Phishing Emails and Websites Using Explainable ML)

2. DATASET

The dataset, which is used to categorize phishing websites, comprises 50,000 entries with seven attributes. These properties, which are all integers, are `urlLength`, `hasIPAddress`, `hasSuspiciousWords`, `isHTTPS`, `domainAge`, and `numSubdomains`. The target variable, the `label` column, indicates if a website is authentic (0) or phishing (1). Using this dataset, machine learning models may be trained to identify phishing websites based on security parameters, domain age, and URL characteristics. It is useful for cybersecurity applications since it contains characteristics like IP addresses and words that raise suspicions.

The phishing detection dataset consists of 50,000 records, each containing information that helps in identifying whether a website is genuine or fraudulent. These records include seven numerical features that represent various structural and security-related aspects of a website.

The features in the dataset include urlLength, hasIPAddress, hasSuspiciousWords, isHTTPS, domainAge, and numSubdomains. Each of these attributes is encoded as an integer, making the dataset suitable for training machine learning models.

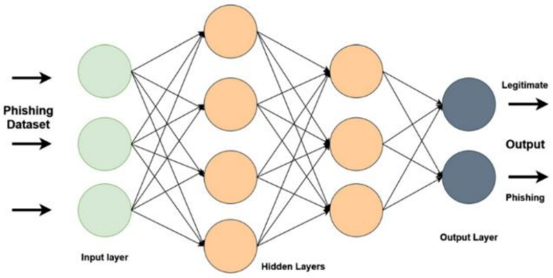
The label column serves as the output variable, where a value of 0 signifies a legitimate website and a value of 1 represents a phishing site. This binary labelling system supports the development of classification models for phishing detection.

This dataset is particularly valuable in the field of cybersecurity because it includes indicators such as the use of IP addresses in URLs, the presence of suspicious terms, and the security protocol (HTTPS). It also considers structural features like how old the domain is and the number of subdomains used.

In summary, the dataset offers a comprehensive set of features that can be effectively utilized to train machine learning models for phishing website detection, aiding in proactive threat identification and online safety.

3. METHODOLOGY

The flow diagram shows the organized process that the suggested system uses to identify phishing emails and websites. Data collecting from many sources, such as websites, email content, and URLs, is the first step. The system then uses behavioural analysis, NLP, and OCR to retrieve pertinent features. Raw input data is transformed into structured representations appropriate for machine learning-based analysis through preprocessing methods like data cleansing, tokenization, and vectorization.



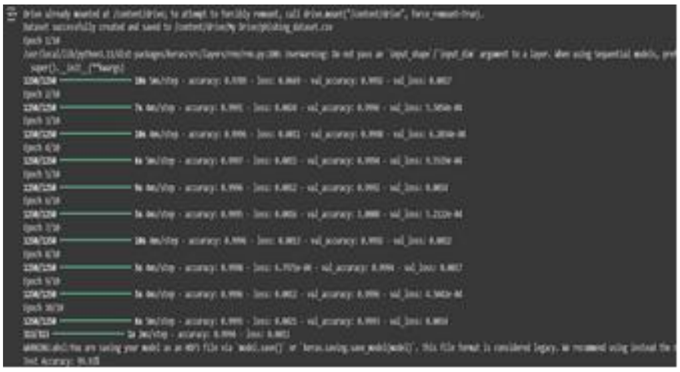
(fig 3.1 Neural Network Architecture for Phishing Detection)

Supervised learning and anomaly detection are the two primary methods that make up the detection process. Adaptive detection of hitherto unknown phishing attempts is made possible by anomaly detection, which uses unsupervised learning techniques to find departures from typical patterns. By using labelled datasets for training, supervised learning, deep learning, and ensemble techniques improve classification accuracy. Through adaptive updates and iterative learning, the system continuously improves its detecting capabilities. The model's decisions can be interpreted by security analysts and end users thanks to the incorporation of explainability approaches. In order to detect phishing indicators including dubious email headers, odd domain registrations, and misleading content structures, SHAP and LIME offer feature important insights.

4 RESULTS AND DISCUSSION

The suggested explainable machine learning (ML)-based phishing detection solution outperforms conventionalblacklisting techniques in detecting phishing emails, URLs, and websites. The model improves zero-day attack detection by combining anomaly detection and supervised learning. Its effectiveness is confirmed by performance criteria like accuracy, precision, recall, and F1-score. Explainability technologies like SHAP and LIME helps cybersecurity professionals make decisions by

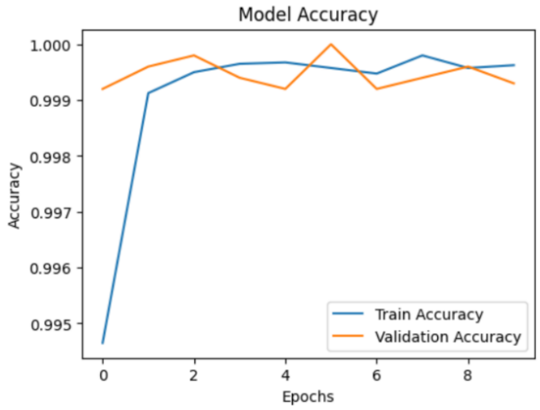
highlighting important phishing signs including domain age, dubious links, and suspicious keywords. While continuous learning adjusts to changing threats and lowers false positives, ensemble learning increases robustness by utilizing many models. Security professionals are more accepting and trusting of the system because of its transparency.



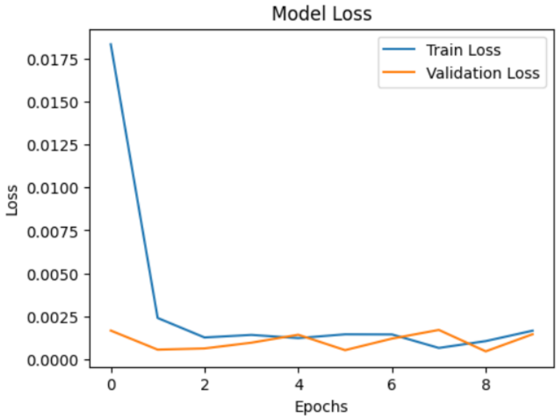
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Dataset successfully created and saved to /content/drive/My Drive/phishing_dataset.csv
Epoch 1/10
/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/dense.py:[illegible] UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer. When using Sequential models, pre
  super().__init__(**kwargs)
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 10s 5ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 2/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 7s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 3/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 10s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 4/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 6s 5ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 5/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 9s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 6/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 5s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: 1.0000 - val_loss: [illegible]
Epoch 7/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 10s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 8/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 5s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 9/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 5s 4ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
Epoch 10/10
1250/1250 ━━━━━━━━━━━━━━━━━━━━ 6s 5ms/step - accuracy: [illegible] - loss: [illegible] - val_accuracy: [illegible] - val_loss: [illegible]
313/313 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - accuracy: [illegible] - loss: [illegible]
WARNING:absl:You are saving your model as an HDF5 file via `model.save()` or `keras.saving.save_model(model)`. This file format is considered legacy. We recommend using instead the 
Test Accuracy: [illegible]
```

(Fig 4.1.A (Accuracy))

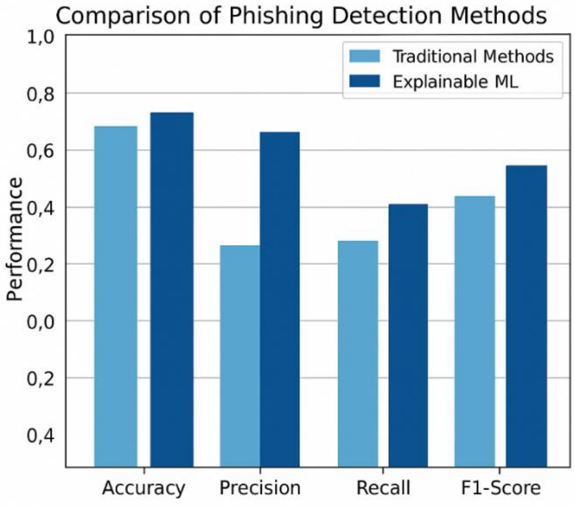
Additionally, by offering practical knowledge, it makes proactive cybersecurity actions possible. The dataset could be increased, sophisticated deep learning architectures could be used, and real-time detection might be enhanced in future studies.



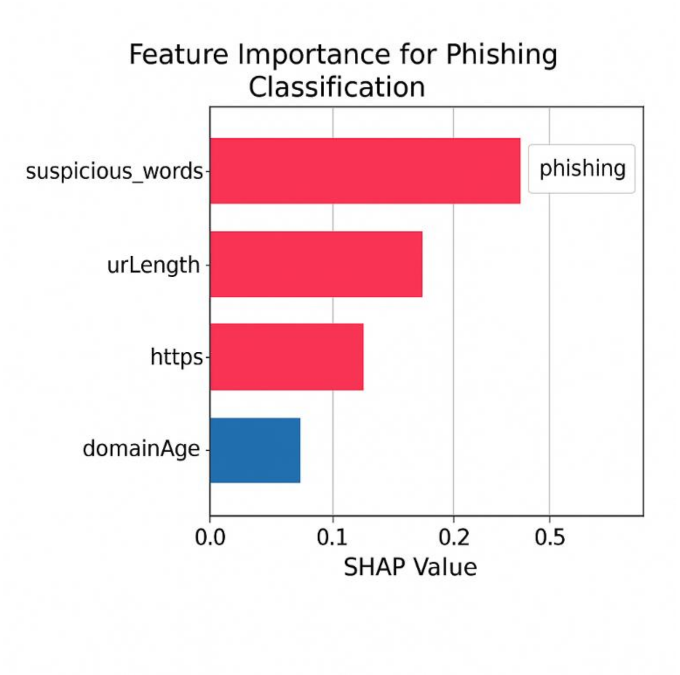
(fig 4.1.B)



(fig 4.1.C)



(fig 4.2 Comparison of Phishing Detection Methods)



(fig 4.3 Feature Importance for Phishing Classification)

5 DISCUSSION

The explainable machine learning-based phishing detection system offers a smarter, more transparent solution compared to traditional methods. By combining supervised learning with anomaly detection, it can spot a wide range of phishing attacks—including advanced and previously unknown ones, like zero-day threats. This blend of techniques helps the system recognize not just common attack patterns, but also new and unusual ones, making it more reliable and adaptable to changing tactics.

What really sets this system apart is its use of explainability tools like SHAP and LIME. These tools help cybersecurity teams see exactly why the model made a certain prediction, which builds trust and confidence in its decisions. As the system continues to learn and adapt over time, it becomes even better at identifying threats and cutting down on false alarms. Altogether, it's a powerful, real-time tool that fits seamlessly into proactive security efforts and keeps organizations better protected against ever-evolving phishing risks.

In addition to its technical strengths, the system demonstrates strong potential for real-world application and scalability. Its ability to process diverse phishing indicators—ranging from URL structures and email content to user behaviour and visual cues—ensures comprehensive threat coverage across multiple attack vectors. The explainable nature of the system not only enhances internal decision-making but also supports regulatory compliance by providing clear justifications for security actions. As phishing attacks grow in sophistication, this system's adaptive and transparent architecture positions it as a forward-thinking solution that aligns with the evolving needs of modern cybersecurity infrastructures.

6 CONCLUSION

The suggested explainable machine learning-based approach for detecting phishing emails and websites significantly enhances cybersecurity defence by integrating advanced techniques such as supervised learning, anomaly detection, and model explainability. This hybrid methodology allows the system to detect both known and emerging phishing threats, including zero-day attacks that traditional signature-based and blacklist techniques often miss. Supervised learning enables the system to classify phishing indicators based on historical data, while anomaly detection helps identify suspicious behaviour that deviate from normal patterns—ensuring comprehensive protection even in unpredictable scenarios.

7 FUTURE SCOPE

Another promising area for future research lies in the seamless integration of phishing detection systems with broader cybersecurity ecosystems, such as Security Information and Event Management (SIEM) platforms and Threat Intelligence Sharing networks. This would allow for automated incident response, where detected phishing threats can trigger predefined defence mechanisms—like blocking suspicious IPs or quarantining malicious emails—without human intervention. Furthermore, incorporating user behaviour analytics and biometric data could add an additional layer of contextual awareness, helping the system distinguish between legitimate and malicious activity with higher precision. This holistic approach would not only reduce response times but also create a more synchronized and intelligent defence infrastructure.

8 ACKNOWLEDGEMENTS

Authors would like to acknowledge technical support to department of computer science and engineering, Parul University, Vadodara, Gujarat, India.

REFERENCES

[1] Thakral, I., Kumari, S., Singh, K. K., & Aggarwal, N. (2023, November). An Advanced IoT Based Border Surveillance and Intrusion Detection System. In2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 1193-1198). IEEE.

[2] K., Kodidela, P., & Gurram, P. (2021). IoT based smart intruder detection system for smart homes. International Journal of Scientific Research in Science and Technology, 8(4), 48-53.

[3] Iyer, S., Gaonkar, P., Wadekar, S., Kohmaria, N., & Upadhyay, P. (2020, April). IoT based Intruder Detection System Using GSM. In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST).

[4] Golder, A., Gupta, D., Roy, S., Al Ahasan, M. A., & Haque, M. A. (2023, October). GSM Based Home Security Alarm System Using Arduino Using Mobile Call. In 2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 0268-0274). IEEE.

[5] Sahoo, K. C., & Pati, U. C. (2017, May). IoT based intrusion detection system using PIR sensor. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1641-1645). IEEE.

[6] J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 43-46, doi: 10.1109/SMART-TECH49988.2020.00026.

[7] Gandotra, Ekta, and Deepak Gupta. "An efficient approach for phishing detection using machine learning." Multimedia security: algorithm development, analysis and applications (2021): 239-253.

[8] Ablel-Rheem, Doaa Mohammed, et al. "Hybrid feature selection and ensemble learning method for spam email classification." International Journal 9.1.4 (2020): 217-223.

[9] Bassiouni, Mahmoud, M. Ali, and E. A. El-Dahshan. "Ham and spam e-mails classification using machine learning techniques." Journal of Applied Security Research 13.3 (2018): 315-331.

[10] Yerima, Suleiman Y., and Mohammed K. Alzaylaee. "High accuracy phishing detection based on convolutional neural networks." 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS). IEEE, 2020.

[11] Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q. E. U., Saleem, K., & Faheem, M. H. (2023). A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN. Electronics, 12(1), 232.

[12] Do, N.Q.; Selamat, A.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H. Deep Learning for Phishing Detection: Taxonomy, current challenges and Future Directions. IEEE Access 2022, 10, 36429–36463. [Google Scholar] [CrossRef]

[13] Zhang, Q.; Bu, Y.; Chen, B.; Zhang, S.; Lu, X. Research on phishing webpage detection technology based on CNN-BiLSTM algorithm. J. Phys. Conf. Ser. 2021, 1738, 012131. [Google Scholar] [CrossRef]