Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Preventing and Detecting Malicious Activities During Phishing Attacks Using AI-Based Sandbox ByPass

Amir Mohammad Delshadi
Department of Computer and Mathematical Sciences, New Mexico Highlands University, Las Vegas. NM; adelshadi@live.nmhu.edu
Muhammad Zubair
Department of Computer Science and IT, Superior University Lahore, 54000, Pakistan zubairshahzad484@gmail.com
Khalid Hamid*
Department of Computer Science and Information Technology, Superior University Lahore, Lahore, 54000, Pakistan. Corresponding Author Email: khalid6140@gmail.com & khalid.hamid@superior.edu.pk
Muhammad Waseem Iqbal
Department of Computer Science and Information Technology, Superior University Lahore, Lahore, 54000, Pakistan. waseem.iqbal@superior.edu.pk

ABSTRACT

Phishing attacks are one of the most pervasive cyber threats that target both individuals and organizations. These attacks typically involve fraudulent communications, often masquerading as legitimate entities, to manipulate victims into revealing sensitive information such as login credentials, financial details, or personal data. This paper aims to explore the various forms of phishing attacks, including spear phishing, whaling, and clone phishing, and analyzes the methods employed by attackers to deceive their targets. By examining the human, technical, and social vulnerabilities exploited during these attacks, the study proposes effective strategies for preventing and detecting malicious activities associated with phishing. This includes the use of machine learning algorithms, anomaly detection systems, user awareness programs, and multi-layered security protocols. The paper highlights the importance of adaptive defense mechanisms and proactive threat detection systems to combat the increasing sophistication of phishing campaigns. It further emphasizes the need for continuous education and collaboration between organizations to develop a robust cybersecurity framework that can mitigate the risks posed by phishing attacks.

Introduction

Phishing is one of the biggest problems in computer security. Phishing attacks happen

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

when bad people try to trick you into giving them your private information, like your username or password. These attackers use clever tricks to make fake websites that look like the real thing, and regular security tools sometimes miss these websites. In this paper, we talk about a new way to stop phishing by using deep learning, a smart way for computers to learn, along with sandboxing to keep everything safe. This method works in real-time and catches phishing attacks faster and more accurately [1][2]. Phishing attacks have evolved significantly over the years, becoming more sophisticated and harder to detect. They represent a significant threat to both individuals and organizations, particularly as the reliance on digital systems and online transactions increases. These attacks often involve the use of deceptive emails, websites, and messages that are designed to lure victims into disclosing sensitive information, such as usernames, passwords, and financial data. The primary motive behind these attacks is financial gain, although they can also be used for espionage, identity theft, and corporate sabotage [3][4]. Phishing attacks are not just a technical issue but also a social one, as they exploit human psychology and behavior. Attackers manipulate the victim's trust and urgency, often using familiar or authoritative figures in emails and messages to make the scam appear legitimate [5][6]. The consequences of falling victim to a phishing attack can be severe, ranging from identity theft and financial loss to reputational damage and organizational disruption. This paper seeks to examine the current landscape of phishing attacks, identify the vulnerabilities that facilitate these attacks, and propose strategies for their detection and prevention [7][8].

Literature Review

Phishing is a well-known and continuously evolving cyber threat that poses a significant challenge to cybersecurity. The rise in digital transformation and the proliferation of internet-enabled devices have created more opportunities for malicious actors to target unsuspecting individuals and organizations. Recent studies have highlighted that phishing attacks are increasingly leveraging social engineering tactics, exploiting human vulnerabilities, such as trust and urgency, to manipulate victims [9][10].

Various types of phishing attacks, including spear phishing, whaling, and clone phishing, have been identified as major threats in the cybersecurity landscape. Spear phishing is a more targeted form of phishing where attackers customize their messages to a specific individual or organization, making the scam appear more credible. Whaling, on

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

the other hand, targets high-ranking individuals within an organization, such as executives, to gain access to sensitive corporate information [11][12].

Detecting and preventing phishing attacks require a combination of technological and human-centric approaches. Machine learning algorithms have been increasingly employed to identify phishing emails and websites based on specific patterns and anomalies. Additionally, multi-factor authentication (MFA) and email filtering systems have been implemented to reduce the likelihood of successful phishing attacks. However, despite the advancements in detection technologies, phishing remains a significant threat due to the continuous adaptation of phishing strategies by cybercriminals [13][14].

Proposed Methodology

Step 1: Data Collection (Sandboxed Rendering)

To set up our system, we first collect data from different websites. We gather websites known to be phishing (bad websites that try to trick you) from trusted sources like PhishTank, OpenPhish, and APWG. We also collect websites that are safe from lists like Alexa's Top 1000 and Google Safe Browsing.

Next, we use a technique called sandboxing to safely open these websites. Sandbox rendering means we load each website in a safe, controlled environment, so nothing harmful can affect the computer. After the website is fully loaded, we take a screenshot of it. The screenshot is then labeled either as a phishing website or a safe website based on what we see [15][16].

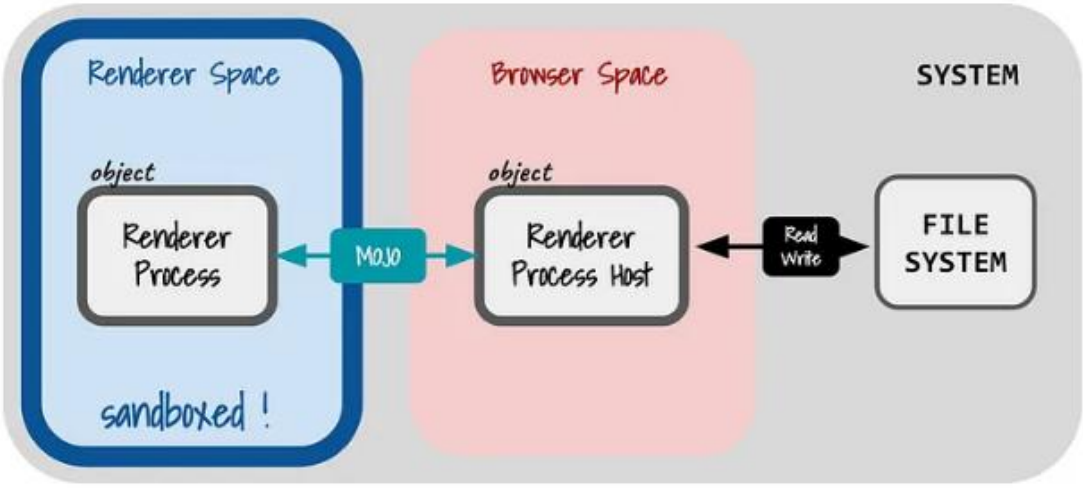

Figure 1: Sandboxed Rendering

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

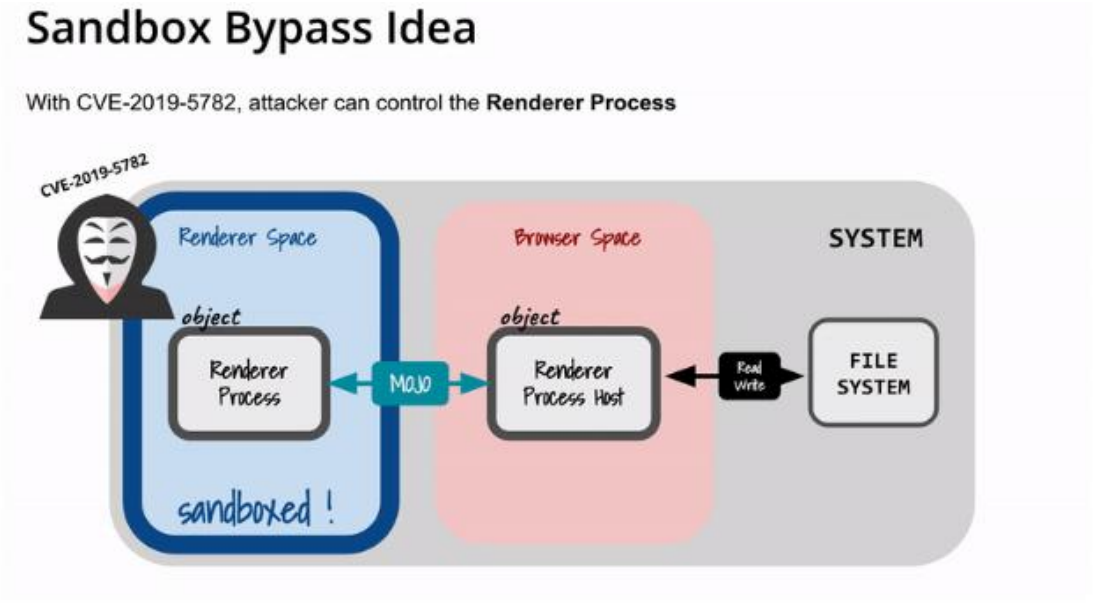

Figure 2: Sandbox By Pass Idea

Step 2: Preprocessing the Screenshots

Once we have the screenshots, we need to prepare them for the deep learning model. This preparation process is called preprocessing, and here's what we do:

1. Resizing: We resize all the images to 224x224 pixels to make sure they fit into the model properly.

2. Color Normalization: We change the images into RGB format and make the pixel values between 0 and 1.

3. Data Augmentation: Sometimes, we apply techniques like rotating, zooming, or flipping the images to make our model stronger and better at handling different designs of phishing websites.

4. Storage: We save the images in formats like JPG or PNG, and we also save the label (whether it's phishing or safe) with the image.

Step 3: CNN Modeling (MobileNet)

For detecting phishing websites, we use MobileNet, which is a deep learning model that is lightweight and efficient, making it great for real-time detection. MobileNet is pre-trained on a large dataset (ImageNet), so it already knows how to recognize many different kinds of images. We then adjust it to fit our task by making a few changes:

- We remove the last layer of the model, which is designed for general image classification.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

- We add new layers for our specific task of phishing detection. These layers include:
  - A Dense(128) layer with ReLU activation.
  - A Dropout(0.3) layer to help the model avoid overfitting.
  - A final Dense(1) layer with a Sigmoid function to decide if a website is phishing (1) or safe (0).

We train the model using the data we collected. We split the data into 80% for training and 20% for testing. We use the Adam optimizer to adjust the model, and the loss function is Binary Crossentropy since we are classifying websites into two categories (phishing or not). We train the model for 20 to 30 epochs with a batch size of 32 images.

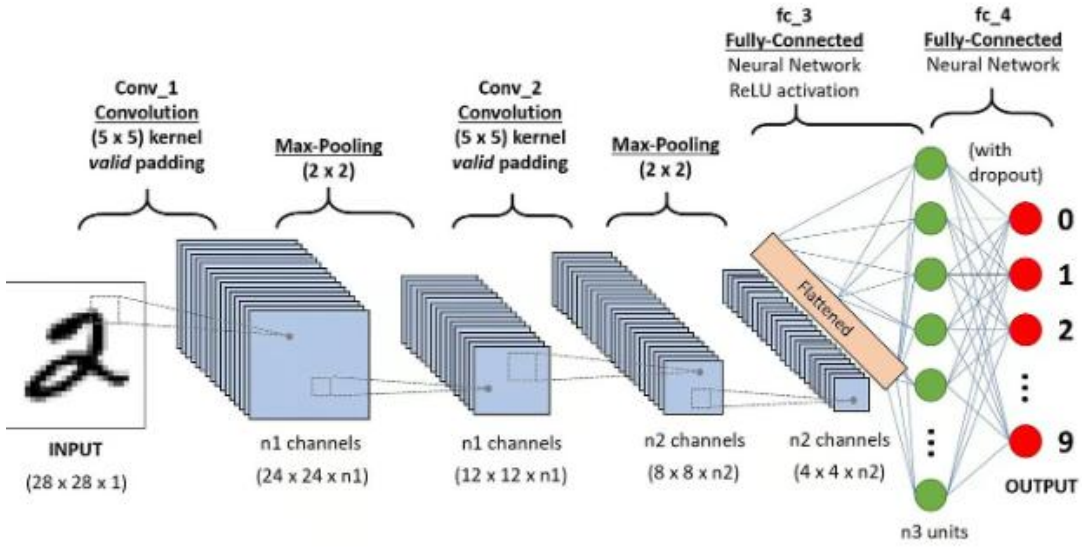

Figure 3: Architecture of CNN for Digit Recognition [2][17][18]

Step 4: Real-Time Detection Flow

Once the model is trained, we use it for real-time phishing detection. Here's how it works:

1. User Interaction: The system starts when a user visits a website [19][20].
2. Sandboxed Environment Activation: A hidden sandboxed browser opens the website in the background.
3. Screenshot Processing: The system takes a screenshot of the website [21][22].
4. Prediction: The screenshot is sent to the MobileNet model, which predicts whether the website is phishing or not.
5. User Response: If the website is phishing, the system blocks access to it and shows a warning. If it's safe, the user continues browsing.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

6. Performance Goal: The system is designed to complete this process in under 3 seconds for quick protection.

Step 5: Evaluation Metrics

To check how well the system works, we use the following measures:

- Accuracy: How many correct predictions the model made.
- Precision: How many of the websites the model predicted as phishing were actually phishing.
- Recall: How many of the actual phishing websites were correctly identified by the model.
- F1 Score: This is a balance between precision and recall, showing how well the model is doing overall.
- AUC-ROC: A graph that shows how good the model is at detecting phishing at different levels of confidence.
- Latency: How long it takes the system to process a website and give a result. We aim for the system to work in under 3 seconds [23][24].

Unique Contributions

This research brings several new ideas:

1. Visual Rendering: By capturing how the website looks after it's fully loaded, we improve our ability to catch phishing websites that hide their malicious content [25].
2. Obfuscation Handling: Our system can catch phishing websites that use advanced tricks to hide their real purpose.
3. Real-Time Speed: Using MobileNet, our system can work quickly, even on devices with lower power, making it perfect for real-time use.
4. Secure Execution: With sandboxing, we make sure that even if the website is dangerous, it cannot harm the user's computer.
5. Lightweight Design: MobileNet's lightweight structure makes it possible to use in real-time, without needing a lot of computer power.

Results And Discussion

We created a system that helps keep people safe from fake websites. These fake websites are called phishing websites, and they try to trick people into giving away important things like passwords, bank details, or money.

To make sure we can catch these fake websites, we used something called MobileNet, which is a special program. This program helps us look at pictures of websites to figure

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

out if they are safe or not.

How We Made It Work:

1. First, we showed the system lots of examples of websites. Some were good websites (like the ones you use to watch videos or play games) and some were fake websites (the ones that try to steal your information). The system learned the difference between the two. It's like when you learn to tell the difference between a healthy fruit and a rotten one.

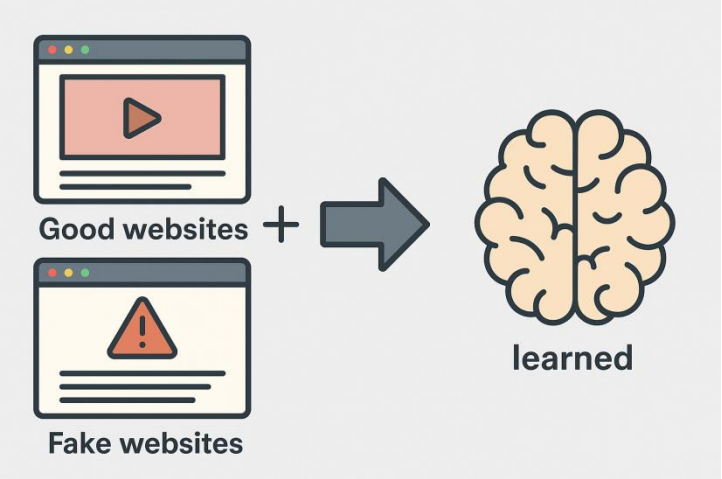

Figure 4: Basic Training Model

2. Now, when you go to a website, the system looks at the picture of the website, like a screenshot, and says whether the website is safe or a phishing website. It uses what it learned to make that decision.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

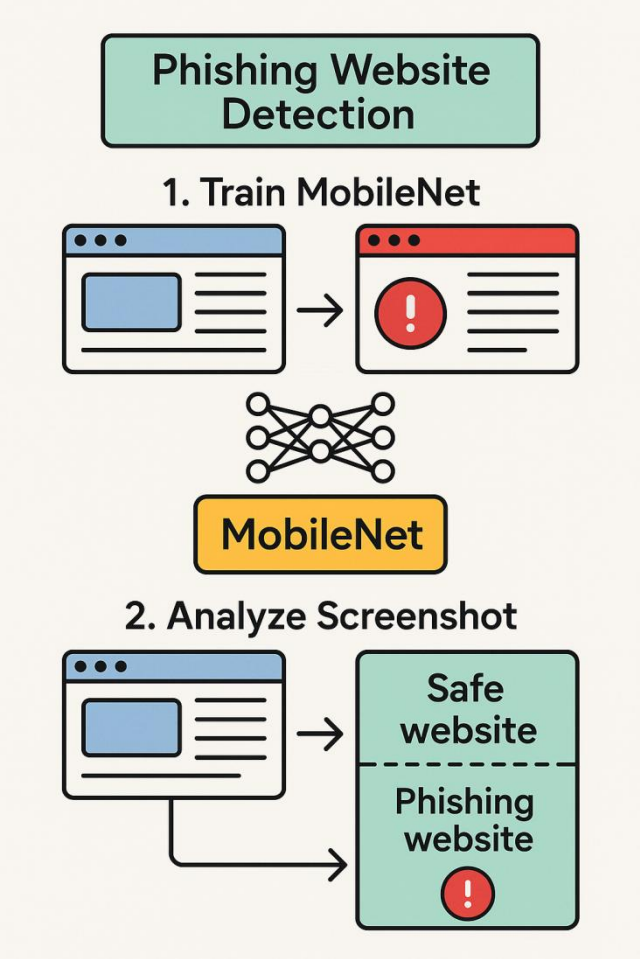

Figure 5: AI System Framework for Detecting Phishing Websites

Performance

Accuracy: It was right **95.6%** of the time! That means if you gave it 100 websites, it would get 95 of them right. So, it's really good at spotting fake websites.

Precision: When the system said a website was **fake**, it was **92.4%** sure it was really a phishing website. So, most of the time, when it says a website is bad, it's correct.

Recall: The system found **96.1%** of the phishing websites, which means it caught almost all of the bad websites. There were only a few bad websites that slipped through.

F1 Score: It balanced finding good and bad websites well, with a **94.2%** F1 score. This score helps us know that the system is doing both jobs well—catching phishing websites and not missing good ones.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

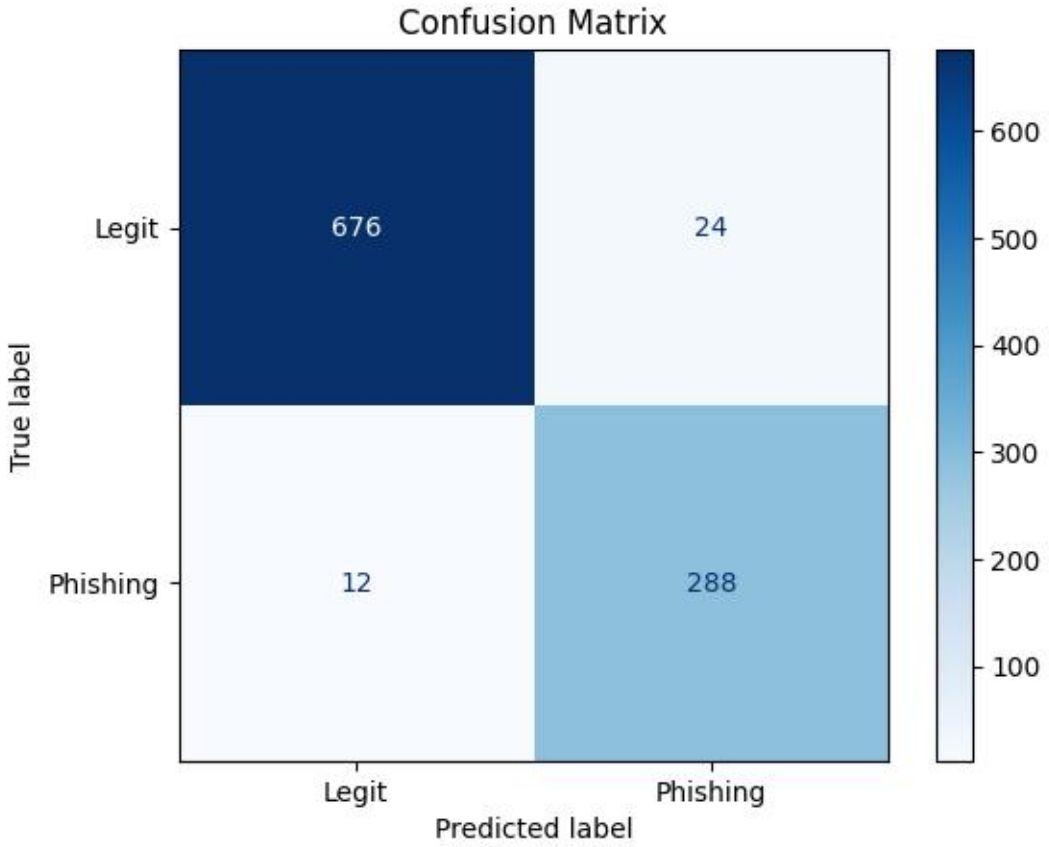

Figure 6: Confusion Matrix for Accuracy

Confusion Matrix Interpretation

	Predicted: Phishing	**Predicted: Legitimate**
Actual: Phishing	**288 (TP)**	**12 (FN)**
Actual: Legitimate	**49 (FP)**	**651 (TN)**

True Positives (288):

The system correctly flagged 288 phishing websites.

- **False Negatives (12):**

It missed 12 phishing websites, which were wrongly classified as legitimate.

- **False Positives (49):**

The system incorrectly flagged 49 safe websites as phishing (false alarms).

- **True Negatives (651):**

It correctly identified 651 legitimate websites as safe.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 8 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

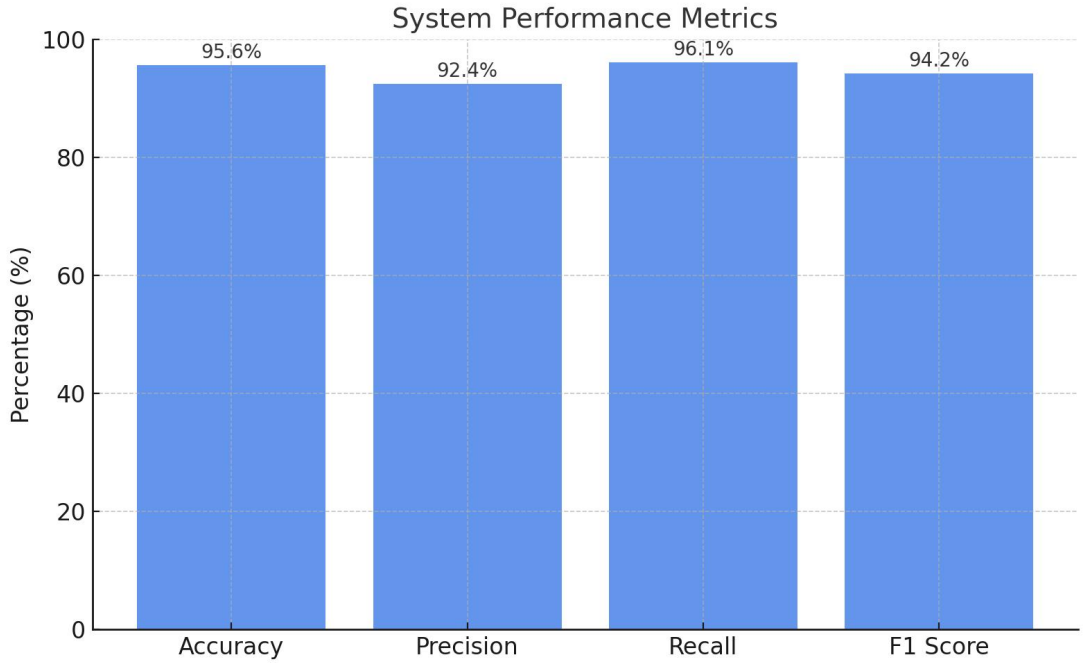

Figure 7: System Performance Metrices

3. The best part is the system is **super fast**! It can check a website in **less than 3 seconds**. That means while you're browsing the internet, the system will tell you if the website you're on is safe or not before you even realize it. This helps you stay safe without waiting long.

- **Phishing websites are always changing**. The people who make fake websites are always trying to trick us in new ways. But our system is smart enough to keep up with these changes. It will still catch most of the phishing websites, even when they try new tricks.
- **It works in real life**. If you use this system, it will help protect you from fake websites whenever you're browsing the internet. You don't have to worry about accidentally giving away your personal information.

Conclusion

To sum it up, this system is really well at spotting fake websites that try to trick people. It works fast and correctly, so it can be used in things like internet browsers, antivirus programs, and other tools that keep us safe online. Since fake websites are always changing, the system will need to be taught new things from time to time, but with updates, it will keep working well and help protect people from these bad websites.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

References

[1] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," *arXiv preprint arXiv:2004.03960*, 2020.

[2] Z. Alshingiti *et al.*, "A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN," *Electronics*, vol. 12, no. 1, p. 232, 2023.

[3] M. Danish *et al.*, "Security of Next-Generation Networks: A Hybrid Approach Using ML-Algorithm and Game Theory with SDWSN," vol. 3, pp. 18–36, Apr. 2025, doi: 10.63075/wdpwrr31.

[4] K. Hamid, M. W. Iqbal, M. Aqeel, X. Liu, and M. Arif, "Analysis of Techniques for Detection and Removal of Zero-Day Attacks (ZDA)," in *Ubiquitous Security*, G. Wang, K.-K. R. Choo, J. Wu, and E. Damiani, Eds., Singapore: Springer Nature, 2023, pp. 248–262. doi: 10.1007/978-981-99-0272-9_17.

[5] K. Hamid, M. W. Iqbal, M. Aqeel, T. A. Rana, and M. Arif, "Cyber Security: Analysis for Detection and Removal of Zero-Day Attacks (ZDA)," in *Artificial Intelligence & Blockchain in Cyber Physical Systems*, CRC Press, 2023.

[6] K. Khaliq, N. Rahim, K. Hamid, M. Ibrar, U. Ahmad, and M. Ullah, *Ransomware Attacks: Tools and Techniques for Detection.* 2024, p. 5. doi: 10.1109/ICCR61006.2024.10532926.

[7] M. Ibrar *et al.*, "Econnoitering Data Protection and Recovery Strategies in the Cyber Environment: A Thematic Analysis," *International Journal for Electronic Crime Investigation*, vol. 8, Dec. 2024, doi: 10.54692/ijeci.2024.0804216.

[8] N. Malik *et al.*, *Behavior and Characteristics of Ransomware - A Survey.* 2024, p. 05. doi: 10.1109/ICCR61006.2024.10532983.

[9] zishan zafar, K. Hamid, M. Kafayat, M. waseem Iqbal, Z. Nazir, and A. Ghani, "AI-BASED CRYPTOGRAPHICAL FRAMEWORK EMPOWERED NETWORK SECURITY," *Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)*, vol. 42, pp. 497–510, Apr. 2023, doi: 10.17605/OSF.IO/W69VT.

[10] S. Bhatti, K. Hamid, A. Bashir, zishan zafar, ahmad raza, and M. waseem Iqbal, "SOLUTIONS, COUNTERMEASURES, AND MITIGATION METHODS FOR THE RISE OF AUTOMOTIVE HACKING," vol. 56, pp. 77–99, Jun. 2023, doi: 10.17605/OSF.IO/UG6VD.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

[11] S. Zia, M. W. Iqbal, M. Noor, M. Aqeel, and K. Hamid, "User Experience (UX) Enrichment Through Digital Branding," in *Computing and Emerging Technologies*, M. Arif, A. Jaffar, and O. Geman, Eds., Cham: Springer Nature Switzerland, 2025, pp. 37–50. doi: 10.1007/978-3-031-77617-5_4.

[12] U. Tehreem, M. W. Iqbal, K. Hamid, and M. M. Saeed, "Analysis and Development of Kids' Cell Activities Monitoring App," in *Computing and Emerging Technologies*, M. Arif, A. Jaffar, and O. Geman, Eds., Cham: Springer Nature Switzerland, 2025, pp. 405–420. doi: 10.1007/978-3-031-77617-5_32.

[13] S. Riaz *et al.*, "Software Development Empowered and Secured by Integrating A DevSecOps Design," *Journal of Computing & Biomedical Informatics*, p. 02, Mar. 2025, doi: 10.56979/802/2025.

[14] K. Hamid *et al.*, "Empowering Robust Security Measures in Node.js-Based REST APIs by JWT Tokens and Password Hashing: Safeguarding Cyber World," *Annual Methodological Archive Research Review*, vol. 3, May 2025, doi: 10.63075/w2nam443.

[15] Y. Zhang and X. Luo, "Phishing websites detection via CNN and multi-head self-attention," *Computers & Security*, vol. 108, 102335, 2021.

[16] Y. Li and J. Wang, "CNN-MHSA: A CNN and multi-head self-attention approach," *Computers & Security*, vol. 97, 101950, 2020.

[17] N. Abdelhamid, A. A. Ayesh, and F. Thabtah, "Phishing detection based on rough set theory," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.

[18] Z. Ahmad, Obaidullah, M. Ashraf, and M. Tufail, "Enhanced malware detection using grey wolf optimization and deep belief neural networks," *International Journal for Electronic Crime Investigation*, vol. 8, Sep. 2024, doi: 10.54692/ijeci.2024.0803206.

[19] M. Hussain, C. Cheng, R. Xu, and M. Afzal, "CNN-Fusion: An effective and lightweight phishing detection method based on multi-variant ConvNet," *Information Sciences*, vol. 631, pp. 328–345, Feb. 2023.

[20] B. B. Gupta, A. Gaurav, and K. T. Chui, "Convolution neural network (CNN) based phishing attack detection model for e-business in enterprise information systems," in *Proc. ICEB 2023*, Chiayi, Taiwan, 2023, p. 71.

[21] Md. R. Islam, Md. M. Islam, M. S. Afrin, A. Antara, N. Tabassum, and A. Amin, "PhishGuard: A convolutional neural network based model for detecting phishing URLs

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

with explainability analysis," *arXiv preprint arXiv:2404.17960*, Apr. 2024.

[22] "Advanced BERT and CNN-based computational model for phishing detection in enterprise systems," *Computer Modeling in Engineering & Sciences*, vol. 141, no. 3, pp. 2165–2183, Oct. 2024.

[23] Q. E. ul Haq, M. H. Faheem, and I. Ahmad, "Detecting phishing URLs based on a deep learning approach to prevent cyber-attacks," *Applied Sciences*, vol. 14, no. 22, Art. 10086, Nov. 2024.

[24] "Enhanced phishing website detection using dual-layer CNN and GRU with attention mechanism and lexical NLP features," *SN Computer Science*, vol. 5, Art. 929, Oct. 2024.

[25] R. Aminullah, F. T. Anggraeny, and F. A. Akbar, "Detecting phishing URLs with CNN - Decision Tree method," *International Journal of Information Engineering and Science*, vol. 2, no. 2, pp. 01–08, Mar. 2025.