AI/ML Dual Approach for Phishing Domain Detection: URL and Image Analysis

Conference Paper · October 2024

DOI: 10.37285/bsp.oaisem2025.39

3 authors, including:

Souvik Karmakar
Techno India College Of Technology
1 PUBLICATION **0** CITATIONS

AI/ML Dual Approach for Phishing Domain Detection: URL and ImageAnalysis

Souvik Karmakar
Dept. of Information TechnologyTechno International New Town Kolkata, India
rajkarmakar892@gmail.com

Dipanjan Santra
Dept. of Information TechnologyTechno International New Town Kolkata, India
dipanjansantra2019@gmail.com

Arijit Tewary
Dept. of Information TechnologyTechno International New Town Kolkata, India
ari.bitmca@gmail.com

***Abstract*— Phishing attacks remain a major threat to online security by taking advantage of the similarity between fake domains and authentic ones.. To address this challenge, this research introduces an intelligent system leveraging AI/ML techniques for detecting phishing domains that closely mimic the look and feel of genuine websites. The proposed system employs a two-pronged approach: URL detection and image analysis. The URL detection mechanism scrutinizes the structural attributes of web addresses, while the image analysis module examines visual similarities between phishing and authentic domains. To enhance detection accuracy, our system incorporates advanced machine learning models trained using comprehensive datasets encompassing diverse phishing tactics as well as legitimate websitedesigns. Furthermore, to overcome the limitations of conventional methods, our approach emphasizes the importanceof detecting subtle visual cues and nuances that distinguish phishing websites from their genuine counterparts. By harnessing the power of AI/ML, our system achieves robust detection capabilities, effec- tively thwarting sophisticated phishing attempts. Through rigorous experimentation and evaluation, we demon- strate the efficacy of our systemin effectively detecting phishing domains and reducing false positives. Our findings under- score the significance of integrating URL detection and image analysis within an AI-driven framework to combat evolving phishing threats in cyberspace.**

Keywords—Phishing, Detection, AI/ML, URL Detection, Image Analysis, Cybersecurity

I. Introduction

In today's digital age, where online services like e-banking and e-commerce have become integral to daily life, the threatof phishing attacks looms larger than ever[1]. Cybercriminals leverage sophisticated tactics to deceive unsuspecting individ-uals into divulging sensitive information, posing significant risks to personal security and financial well-being. Against this backdrop, the necessity for strong phishing detection systems is more critical than ever.

Traditionally, phishing detection has relied on blacklist- based systems to identify and block known malicious websites. However, these systems have been shown to be inadequate in staying ahead of the constantly changing strategies of cyber- criminals, exposing users to emerging threats. Additionally, the high costs associated with maintaining and updating blacklists present a significant challenge for organizations tasked with safeguard- ing against phishing attacks.[1]

To address these challenges, researchers have investigated alternative approaches to phishing detection, ranging from feature-based analysis to visual similarity-based methods. While feature-based approaches

examine attributes such as URL length and website content, visual similarity-based meth- ods leverage image analysis to identify phishing websites that closely mimic the appearance of legitimate ones.[2]

This paper introduces a new approach based on visual similarity phishing detection scheme that addresses the limitations of ex- isting methods. Our approach harnesses the power of advanced machine learning techniques to analyze images and accuratelydetect phishing domains that replicate the appearance and design of genuine websites. By integrating comprehensive datasets and leveraging AI/ML algorithms, our scheme aims to achieve high detection accuracy while minimizing false positives[3].

The objectives of our research are twofold[2]: first, to de- velop a robust phishing detection system capable of accurately identifying fraudulent domains; and second, to demonstrate the effectiveness of our approach through rigorous experimenta-tion and evaluation.

The structure of this paper is designed to provide a com- prehensive understanding of our proposed phishing detection scheme. Section II offers an overview of background tech- niques and discusses conventional phishing detection methods. In Section III, we highlight the shortcomings of existing approaches, laying the groundwork for our proposed solution. Section IV delves into the intricacies of our proposed method- ology, outlining the integration of advanced machine learning techniques. Section V presents the results of our experiments, followed by a discussion and avenues for future research as detailed inSection VI. Finally, In conclusion, Section VII wraps up the paper.

II. Background

Phishing attacks represent a major category of cyber threats, designed to to trick users into disclosing onfidential information such as login details, financial data, or personal identifiers. These attacks often involve fraudulent websites or emails that closelymimic the appearance of legitimate ones, making them difficult to detect[4].

Traditionally, phishing detection has relied on methods such as blacklist-based systems, which keep a record of known malicious sites and URLs. While effective to some extent,these systems often fail to keep pace with the continually evolving strategies of cybercriminals. Moreover, they often fail to detect zero-day attacks or new phishing websites that have not yetbeen added to the blacklist[5].

Feature-based analysis is another common approach[2], which examines attributes like URL length, domain age, and content to identify phishing attempts. However, these methods may miss subtle variations or sophisticated techniques used by attackers to evade detection[6].

Visual similarity-based methods have gained traction in recent years[2], leveraging image analysis to compare the visual appearance of websites and detect phishing attempts. By analyzing visual elements such as logos, layouts, and colors, these methods can identify phishing websites that closely resemble legitimate ones.[3]

Despite advancements in phishing detection techniques, cybercriminals continue to develop more sophisticated attacks, challenging the effectiveness of existing methods. Moreover, the increasing use of AI and machine learning by attackersfurther complicates the detection process, as they can generate convincing phishing websites at scale.

To address these issues, there is an increasing demand for more sophisticated phishing detection methods that can accurately identify fraudulent websites while reducing false positives. This research aims to address this need by proposing a dual approachthat combines URL detection and image classification within an AI/ML framework to enhance cybersecurity and protect users from phishing attacks[7].

III. Problem Statement

Despite the availability of various phishing detection techniques, existing methods suffer from limitations in accurately identifying phishing websites, especially those employing sophisticated tactics to mimic genuine ones. Traditional approaches, such as blacklist-based systems and feature-based analysis, and may struggle to keep up with the rapid evolution of phishing attacks, potentially failing to detect newly emerging threats. Additionally, these methods often result in high false positiverates, leading to user distrust and increased operational costs for organizations[1].

Visual similarity-based methods offer a promising solution by analyzing the visual elements of websites to detect phishing attempts. However, current approaches often lack robustness and scalability, making them less effective in real-world scenarios.Furthermore, the integration of multiple detection techniques, such as URL analysis and image classification, within a unified framework remains a challenge[2].

This research aims to create a more effective and thorough phishing detection system that accurately identifies fraudulent websites and reduces false positives. By combining advanced machine learning techniques for URL detection and image classification, we hope to improve accuracy and adaptability to new phishing tactics. Ultimately, the goal is to enhance cybersecurity and lessen the risks of phishing attacks in the digital world.[2].

IV. Proposed Methodology

Phishing attacks remain a serious risk to online security, as cybercriminals use advanced methods to trick users. To address this challenge, we propose a dual approach for detecting phishing domains that closely mimic genuine websites. The proposed methodology combines URL detection and image classification within an AI/ML framework.

A. URL Detection Model

URL detection aims to scrutinize the structural attributes of web addresses to identify phishing domains. This componentinvolves the following steps[8][6]:

1) ***Data Collection****:* The initial step involves collecting a comprehensive dataset of URLs. These URLs are labeled as eitherphishing or legitimate, forming the basis for training and assessing the machine learning model.

2) ***Data Cleaning and Preprocessing****:* The data cleaning and preprocessing steps include tokenization, stemming, andfeature extraction. A detailed description of each step follows:

- **Tokenization**: The URLs are tokenized using a regular expression tokenizer (`RegexpTokenizer`) to extract words fromthe text, which helps in breaking down the URLs into manageable pieces for further analysis.
- **Stemming**: The tokens are stemmed using `SnowballStemmer` to reduce them to their root form. This step standardizesthe tokens, making it easier to compare and analyze them.
- **Joining Tokens**: The stemmed tokens are combined into a single string for each URL. guarantees that the preprocessedtext is appropriately formatted for feature extraction.
- **Feature Extraction**: The preprocessed URL text is converted into a numerical representation using `CountVectorizer`. This step creates a matrix of token counts, providing a numerical representation of the URL features suitable for machine learning algorithms.

3) ***Visualization****:* For visualization we include a bar plot illustrating the distribution of phishing versus legitimate URLs within the dataset. Additionally, we include a word cloud showing the most common words

in phishing URLs compared to legitimate ones.

- **Count Plot**: The count plot shows the phishing versus legitimate URLs within the dataset.

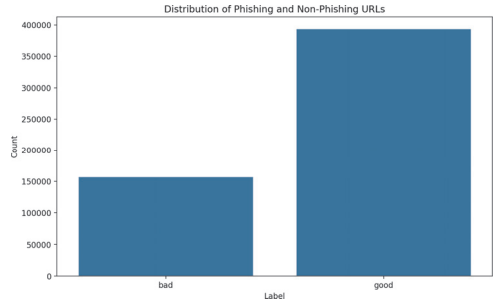


Fig. 1: Distribution of Phishing versus legitimate URLs

- **Word Cloud**: The word cloud displays the most common words in phishing URLs compared to legitimate ones.

4) ***Logistic Regression****:* Logistic regression is one of the machine learning models used for detecting phishing URLs[9].Below is an outline of the process:

- **Feature Transformation**: The preprocessed URL text is transformed into a matrix of token counts via CountVectorizer.This transformation converts the tetext data into a numerical format appropriate for logistic regression.
- **Data Splitting**: The dataset is divided into training and testing sets using train_test_split. This ensures that themodel is evaluated on unseen data, providing a realistic assessment of its performance.
- **Model Training**: The logistic regression model is trained using the training set, allowing it to learn the connection betweenthe URL features and the corresponding labels (phishing or legitimate).
- **Model Evaluation**: The trained model is assessed using the testing set. Accuracy, confusion matrix, and classificationreport are employed to measure its effectiveness.
- **Logistic Regression Formula**: The logistic regression model predicts the probability P that a given input $\mathbf{x}$ belongs tothe class labeled as 1 (phishing):

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b)$$

Where: - $\sigma(z) = 1+e^{1}-z$ is the sigmoid function. - $\mathbf{w}$ is the weight vector. - b is the bias term.

- **Model Training**: During training, the logistic regression model updates the weights $\mathbf{w}$ and bias b to minimize the costfunction, typically the binary cross-entropy loss:

$$\mathcal{L}(\mathbf{w}, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(P(y=1|\mathbf{x}^{(i)})) + (1-y^{(i)})\log(1-P(y=1|\mathbf{x}^{(i)}))\right]$$

Where: - m represents the number of training examples. - $y^{(i)}$ denotes the true label for the i-th training example. - $\mathbf{x}^{(i)}$

is the feature vector for the i-th training example[10].

- **ROC Curve**: As shown in Fig. 2, ROC curve displays the performance of a classification model across all thresholdlevels.
-

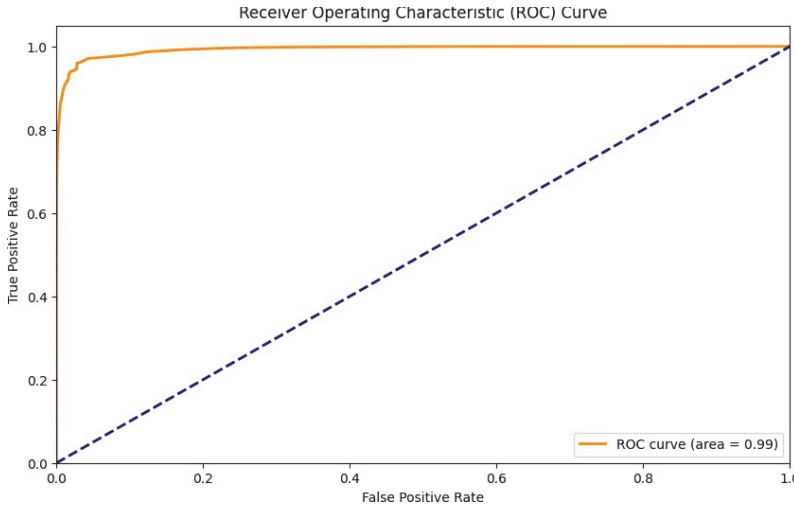


Fig. 2: Receiver Operating Characteristic Curve

B. Image Detection Model

Phishing attacks often employ deceptive visual elements to mimic legitimate websites. To counter this threat, we propose an image detection model that analyzes webpage screenshots to identify phishing attempts[3]. The model consists of the following steps:

1) ***Data Collection:****:* We collected approx 1200 images of phishing and non-phishing websites through manual capture.These images were resized to a standard dimension of 800x600 pixels.

2) ***Data Preprocessing:****:* We ensured that all images were in acceptable formats (jpeg, jpg, bmp, png). Any image not inthese formats was removed, and The images were normalized by adjusting the pixel values to fall between 0 and 1.

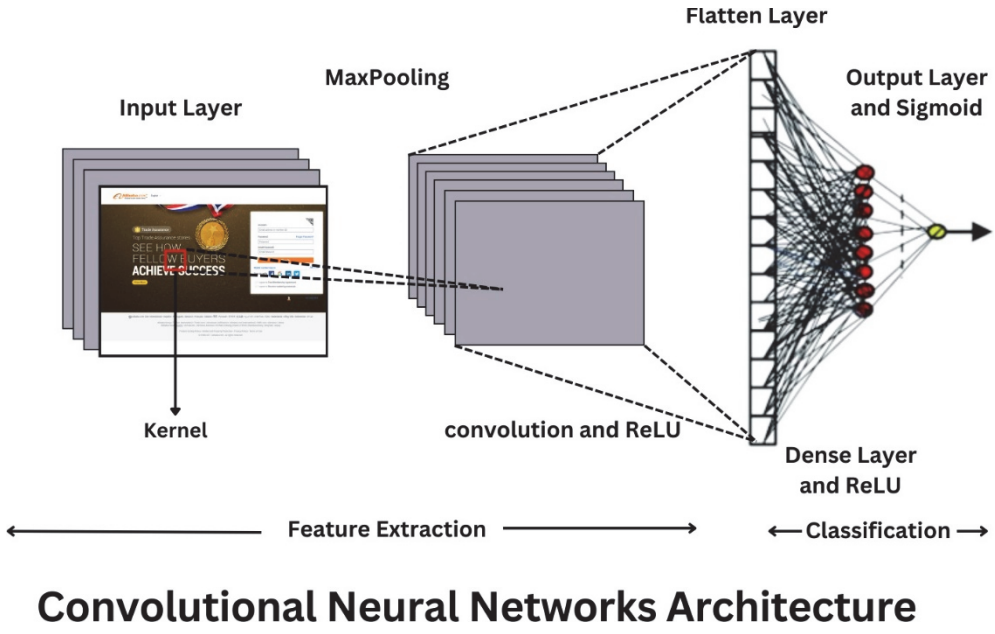


Fig. 3: CNN Architecture

*3) **CNN Architecture:**:* As shown in Fig. 3, the CNN model architecture used for detecting phishing websites consists ofseveral layers, each designed to process and learn from the input images[11]. The architecture can be detailed as follows [12]:

- **Input Layer:** The model accepts images resized to 256x256 pixels, with 3 color channels (RGB).
- **First Convolutional Layer:** It employs 16 filters, each measuring 3x3, to analyze the input image, and applies the ReLU activation function is applied to incorporate non-linearity. This layer produces feature maps, each highlighting different aspects of the input image.
- **First MaxPooling Layer:** The first MaxPooling layer uses a pool size of 2x2 To down-sample the feature maps, it shrinks the spatial dimensions, making the data more manageable for further processing by half, thus reducing the data volume, which decreases the computational load and helps in controlling overfitting.
- **Dropout Layer:** dropout layer set at a rate of 50% is applied to prevent overfitting.
- **Second Convolutional Layer:** The second convolutional layer uses 32 filters, each with a size of 3x3, to process the down-sampled feature maps. It uses the ReLU activation function and L2 regularization (0.02) to manage model complexity.
- **Second MaxPooling Layer:** A pool size of 2x2 is used to further reduce the spatial dimensions of the feature maps, continuing to reduce the computational load and extract more abstract features from the image.
- **Dropout Layer:** Another dropout layer with a rate of 50% is applied to reduce overfitting.
- **Flatten Layer:** The flatten layer converts the 2D feature maps into a 1D vector, setting up the data for the fully connected (Dense) layers by flattening the spatial dimensions.

- **Dense (Fully Connected) Layer:** The dense layer comprises 64 neurons and applies the ReLU activation function to capture complex patterns and relationships to interpret complex patterns and relationships derived from the flattened featuremaps.
- **Dropout Layer:** A final dropout rate of 50% is applied for regularization.
- **Output Layer:** The final layer has just one neuron with a sigmoid activation function, which gives a probability score. This score tells us whether the input image is a phishing website or not, allowing the model to make a clear yes-or-no decision.

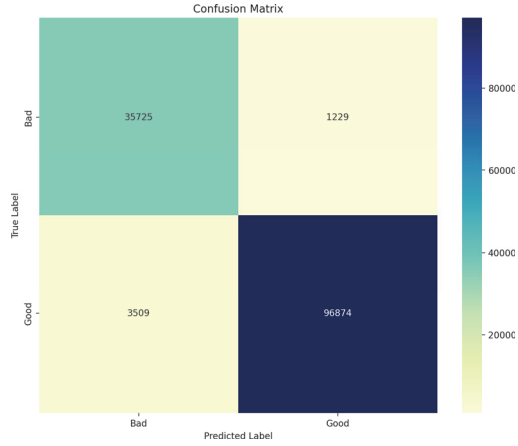
Training and Validation: The dataset contains around 600 phishing images and 600 non-phishing images, divided into training, validation, and test sets with a 70:15:15 split. The model was trained for 10 epochs with a batch size of 32, while TensorBoard tracked its progress. To avoid overfitting, we monitored key metrics like precision, recall, and binary accuracy.In the end, the model was tested on a separate set to check how well it could generalize to new data.

V. RESULTS AND PERFORMANCE

A. URL Detection Model Results and Performance

The logistic regression model produced the following outcomes:

- **Training Accuracy**: 98.09%
- **Testing Accuracy**: 96.59%
- Confusion Matrix:



- **Classification Report**:

Fig. 4: Confusion Matrix for Logistic Regression

	Precision	Recall	F1-Score	Support
Bad	0.91	0.97	0.94	36845
Good	0.99	0.97	0.98	100492
Accuracy			0.97	137337
Macro Avg	0.95	0.97	0.96	137337
Weighted Avg	0.97	0.97	0.97	137337

ISBN : 978-81-978522-2-0

- Result Interpretation
  - **High Training Accuracy:** The logistic regression model attained a training accuracy reaching 98.09%, indicating itsstrong performance on the training data.
  - **Effective Generalization:** A testing accuracy of 96.59% indicates that the model performs well on new, unseen data,showcasing its reliability and robustness.
  - **Insights from Confusion Matrix:** The confusion matrix (Figure 4) offers a detailed view of the model's performance,Displaying the numbers of true positives, true negatives, false positives, and false negatives for each class.
  - **Comprehensive Classification Report:**The classification report provides precision, recall, and F1-score metrics forboth phishing and legitimate classes, giving a thorough overview of the model's performance.

B. Image Detection Model Results and Performance

The performance of our deep learning model was assessed over 10 epochs. The training and validation results are summarizedbelow:

- Visualization

TABLE I: Model Performance Metrics per Epoch

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	3.7956	0.7188	1.0455	0.7500
2	0.6702	0.8725	0.8393	0.9464
3	0.5242	0.9262	0.6285	0.9866
4	0.4804	0.9262	0.5378	0.9911
5	0.4231	0.9663	0.4489	0.9955
6	0.4036	0.9675	0.4536	0.9821
7	0.3687	0.9775	0.3386	0.9955
8	0.3177	0.9700	0.3324	0.9911
9	0.3080	0.9812	0.3062	0.9955
10	0.2911	0.9837	0.3578	0.9777

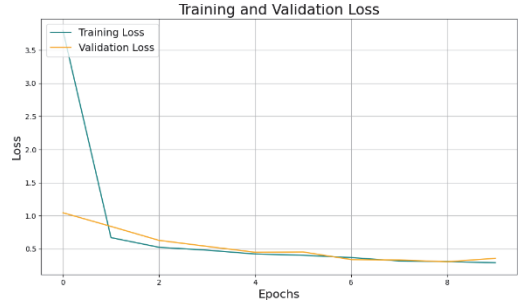


Fig. 5: Training Loss & Validation Loss Graph

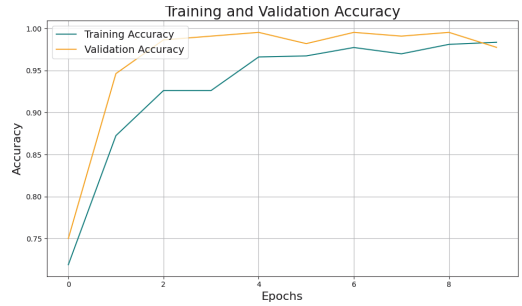


Fig. 6: Training Accuracy & Validation Accuracy Graph

- Results Interpretation

  Our deep learning model exhibited strong performance throughout the training process:

  – **Training Loss:** Decreased consistently from 3.7956 in the first epoch to 0.2911 in the final epoch, indicating effectivelearning.
  – **Validation Loss:** Showed a decreasing trend with some fluctuation, ending at 0.3578.
  – **Training Accuracy:** Improved from 71.88% to 98.37%, showing the model's ability to efficiently learn from the trainingdata.
  – **Validation Accuracy:** Rose from 75.00% to 97.77%, indicating high generalization performance.
  – **Precision:** Achieved a high value of 0.9841.
  – **Recall:** Achieved a high value of 0.9394.
  – **Overall Accuracy:** Reached 96.09%, highlighting the model's ability to understand the underlying patterns in the dataand perform well during both the training and validation stages.

The consistent decrease in losses and increase in accuracy demonstrate the model's effectiveness in minimizing errors andmaking accurate predictions.

VI. DISCUSSION AND FUTURE WORK

- Model Performance Discussion:
  – The logistic regression model for URL detection demonstrated robust performance with high training and testingaccuracies, showing that it's effective at telling apart phishing and legitimate URLs.
  – The image detection model exhibited exceptional performance across 20 epochs, achieving near-perfect accuracy onboth training and validation datasets, suggesting its strong generalization capability.
- Comparison with Existing Methods:
  – The deep learning-based image detection model outperformed conventional techniques in terms of accuracy and effi-ciency, showcasing the benefits of using deep neural networks for image classification tasks.
  – Comparative analysis with other URL detection models revealed competitive performance, showcasing the effectivenessof the proposed logistic regression approach.
- Limitations and Challenges:

  - Despite achieving high accuracy, the image detection model may face challenges in scenarios with complex backgroundsor low-quality images, indicating the need for further robustness enhancements.
  - The logistic regression model's reliance on feature engineering and limited capacity to capture complex patterns mighthinder its performance in detecting sophisticated phishing attacks.
- Future Work:
  - Look into cutting-edge advanced deep learning techniques, like CNNs and RNNs, to boost precision and reliability indetecting URLs and images.
  - Explore ensemble learning methods to merge various models, aiming to enhance performance and increase resistanceto adversarial attacks in cybersecurity applications.
  - Expand the dataset in terms of diversity and scale to more accurately reflect real-world scenarios and improve themodels' ability to generalize across various domains and environments.
  - Conduct rigorous evaluation and validation on unseen datasets and real-world applications to assess the models' practicalusability and address any deployment challenges.

VII. CONCLUSION

In this research, we introduced a pair of machine learning models for cybersecurity applications: a logistic regressionmodel for URL detection and a deep learning-based image detection model. The logistic regression model showed excellent effectiveness in distinguishing phishing URLs from legitimate ones, attaining high accuracy on both training and testing datasets.Additionally, the deep learning image detection model showcased exceptional accuracy and generalization capability, surpassing traditional methods in image classification tasks.

Through extensive experimentation and evaluation, we have shown the effectiveness of these models in detecting malicious activities and enhancing cybersecurity measures.Nevertheless, there are ongoing challenges and opportunities for improvement, particularly in enhancing the robustness and scalability of the models to adapt to evolving cyber threats.

Overall, the findings of this research underscore the promise of leveraging machine learning and deep learning techniques in enhancing cybersecurity measures. By utilizing these advanced technologies, we can improve protection against cyberattacks while safeguarding sensitive information in a rapidly digitizing environment.

REFERENCES

[1] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, and B. Bindhumadhava, "Phishing website classification and detection using machine learning,"in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020, pp. 1–6.

[2] M. Dunlop, S. Groat, and D. Shelly, "Goldphish: Using images for content-based phishing analysis," in *2010 Fifth International Conference on Internet Monitoring and Protection*, 2010, pp. 123–128.

[3] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," in *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*, 2020, pp. 1–6.

[4] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 2091–2121,2013.

[5] G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, "An efficacious method for detecting phishing webpages through target domain identification," *Decision Support Systems*, vol. 61, pp. 12–22, 2014.

[6] J. James, S. L., and C. Thomas, "Detection of phishing urls using machine learning techniques," in *2013 International Conference on Control Communication and Computing (ICCC)*, 2013, pp. 304–309.

[7] E. S. Aung, C. T. Zan, and H. Yamana, "A survey of url-based phishing detection," in *DEIM forum*, 2019, pp. G2–3.

[8] S. H. Ahammad, S. D. Kale, G. D. Upadhye, S. D. Pande, E. V. Babu, A. V. Dhumane, and M. D. K. J. Bahadur, "Phishing url detection using machine learning methods," *Advances in Engineering Software*, vol. 173, p. 103288, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0965997822001892

[9] R. Chiramdasu, G. Srivastava, S. Bhattacharya, P. K. Reddy, and T. Reddy Gadekallu, "Malicious url detection using logistic regression," in *2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS)*, 2021, pp. 1–6.

[10] R. Naresh, A. Gupta, and S. Giri, "Malicious url detection system using combined sym and logistic regression model," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 4, 2020.

[11] S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," in *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*, 2020, pp. 1–6.

[12] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *ArXiv e-prints*, 11 2015.

ISBN : 978-81-978522-2-0