An Intelligent Phishing Website Detection System Using CNN with Multi-Head Self Attention

Kavi Ranjani S G[1], Afreena Sathaf M[2], Aashika M[3], Dr. P. Chellammal[4]
[1,2,3,4] *Computer Science and Engineering J J College of Engineering and Technology Trichy India.*

***Abstract*—Phishing websites are an ever-changing and prevalent risk to the area of the cybersecurity system as they allow attackers to collect sensitive user credentials, financial information, and personal data by spoofing the web interface. Detecting using conventional methods, including blacklist-based and rule-based methods, has serve weakness against those related to zero-day and new phishing campaigns. Moreover, the imbalance of classes in the existing datasets remains to lower the performance of the current machine learning and deep learning solutions. In this paper, an intelligent phishing site detection framework has been introduced, employing a Convolutional Neural Network (CNN) and a Multi-Head Self-Attention (MHSA) system, as a set of features to be extracted and patterns to be identified are based on URLs. The CNN part itself self-trains on discriminative features using the raw sequences of URL characters, and the self-attention components selectively enhances the discriminative features of URL patterns in order to achieve better classification accuracy and low false positive probability. In order to alleviate the negative side effects of the class imbalance, a Generative Adversarial Network (GAN) is used to generated realistic samples of phishing URLs to augment the training corpus and improve generalization of the model. Large-scale experiments reveal that the suggested system is more that the suggested system is more accurate, precise, recalls and F1-score as compared to baseline strategies, as well as real-time and zero-day phishing is demonstrated. The framework provides a scalable and adaptive framework that is appropriate to integrate into dynamic adversarial settings.**

***Index Terms*—detection phishing, convolutional neural network, multi-head self-attention, generative adversarial network, deep learning, and cybersecurity.**

I. INTTRODUCTION

The rapid diffusion of internet technologies and online services has created a significant threat of cyber threats, and phishing attacks can be considered as one of the most common and dangerous ones. The phishing sites are

malicious web pages that clone legitimate sites in order to defraud the users to submit confidential information such as username, password, credit cards and bank information. The aftermath of these attacks is great loss of money and privacy of the users globally. In addition, phishing techniques difficult to detect them.[1]

The traditional approaches to detecting phishing, such as blacklist and whitelist, rely on the fact they maintain the database of the already known malicious URLs.[2] The techniques are effective in detecting the known threats, but they cannot detect phishing sites that are developed recently, as well as those that are zero-day. In order to eliminate such drawbacks, other machine learning algorithms like Naïve Bayes, Support Vector Machines, Decision Trees and Random Forests have been employed. Such approaches are expected to enhanced the accuracy of detection, but they highly depend on manual feature engineering, and they are likely to experience high false positive rates and an unbalanced dataset.[3]

In the recent years, deep learning has emerged as a potential technology in automated extraction of features and pattern recognition. Convolutional Neural Networks (CNNs) have shown these encouraging results in phishing detection to be able to acquire discriminative features directly out of raw URLs.[4] The classical CNN models, however, may not prove useful in the long-range dependency and contextual relations within the URLs.[5]

To remove these drawbacks, the attention mechanisms, namely Multi-Head Self-Attention have been implemented to enhance the ability of the model further to focus on the appropriate parts of the input data. In this paper, a hybrid phishing detector will be

introduced, which is based on CNN with Multi-Head Self-Attention to optimize the accuracy of classification.[6]
In addition, a Generative Adversarial Network (GAN) is used to address the issue of class imbalance which is usually characteristics of phishing samples and generate fake phishing URLs. Such a strategy of data augmentation makes the training set more diverse and the model stronger.[7] The suggested system will be applicable in real-time cybersecurity implementations as it will be capable of detecting with a high degree of accuracy and reduce false positives and be capable of detecting zero-day phishing attacks as well.[8]

II. EXISTING SYSTEM

Detection of phishing websites has been broadly researched on both the conventional security system and machine learning-based applications. The most popular one is the blacklist/whitelist-based detection system. In this method, there are a set of URLs that are matched to a set database of identified phishing and legitimate sites.[9] When a URL appears in the blacklist, it gets blocked and otherwise, it is perceived as safe. In spite of the fact that this approach is both effective and easy, if fails to identify recently created or zero-day phishing platforms that are not in the database.[10]

To mitigate these shortcomings, detection methods of machine learning have been proposed. Naïve Bayes, Decision Trees, Random Forest, and Support Vector Machines (SVM) are examples of algorithms that are used to categorize websites using manually obtained features, which include characters, age of the domain, and suspicious keywords.[11] Though these techniques are more accurate in detecting than the conventional ones, they are highly reliant on feature engineering huge labeled datasets. Moreover, they are also likely to underperform in handling a phishing attack of speedy change.[12]

Phishing detection has also been applied to deep learning methods, especially Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) networks. These models have the capability of learning patterns from the URL strings and content without necessarily going through considerable feature extraction processes.[13] These models are, however, mostly computationally complex and may face challenges, especially concerning class imbalance, where the number of phishing examples is vastly outstripped by those of normal examples. In spite of all this being said and all these developments having been made, the current systems are facing a number of crucial challenges, most notably being high false positive rates, their inability to work well with zero-day attacks, their dependence on supervised learning, and their inability to adapt to new phishing schemes.[14] In addition, most of them are not able to deal with the imbalance of datasets, leading to biased forecasts and reduced generalization results. All this shows how imperative it is for a stronger phishing detection mechanism, one related to automatic feature extraction, better perception of context, and data balance, to be put in place.[15]

III. PROPOSED SYSTEM

The proposed system outlines a deep learning-based phishing website detection framework, which can enhance the detection accuracy, robustness, and ability to detect zero-day attacks. The model comprises the combination of a convolutional neural network (CNN), the multi-Head Self-Attention mechanism, and the generative adversarial network (GAN) for data augmentation, overcoming the limitations of conventional detection methods.
The overall architecture of the model is based on the end-to-end learning approach, where the raw URL strings can be directly processed without the need for manually extracting the features. The overall system comprises several stages, including the pre-processing of the URLs, CNN-based feature extraction, attention-based contextual enhancement, classification, and the use of the GAN model for balancing the dataset.

1.A.URL Input and preprocessing

The system will accept URLs in the form of strings either from a labeled dataset during training or user inputs during real-time prediction. The preprocessing module will perform several important operations. They are URL validation, noise removal, tokenization, character-level encoding, and conversion of URLs into fixed-length numerical vectors.
The representation of the input URL will be:

$$U=\{c\_1, c\_2, c\_3, idots, c\_n\}$$

The characters in the URL string are represented by c_i.

2.B.CNN-Based feature Extraction

A Convolutional Neural Network (CNN) is used for the automatic extraction of discriminative features from the encoded representation of the URL. The CNN model employs several filters for the extraction of features related to the structure and lexicon of the URLs, such as the occurrence of suspicious keywords, excessive use of special characters, unusual domain structure, and the use of IP-based URLs.

The mathematical representation of the convolution process is as follows:

h_i=f (W*x_i + b)

]

Here, (W) represents the weights of the filters, (x_i) represents the segments of the input, (b) represents the bias, and (f) represents the activation function, which is ReLU. The feature maps are then enhanced using the attention mechanism.

3.C. Multi-Head Self-Attention Mechanism

To improve contextual understanding, a Multi-Head Self- attention mechanism is added after the CNN layers. The attention mechanism gives different attention weights to different parts of the URL, allowing the model to focus on the most relevant phishing-related patterns.

The attention mechanism is defined as:

Text Attention (Q, K, V) =
Text Soft max left frac QK^T sqrt d_k right V

Where (Q), (K), and (V) are the Query, Key, and Value matrices, respectively, and (d_ k) is the scaling factor.

The multi-head attention operation is defined as:

TextAttention (Q, K, V) =textSoftmax left frac QK^T sqrt d_k right V

Where (Q), (K), and (V) represent the Query, Key and Value matrices, respectively, and (d_k) is the scaling factor.

The multi-head attention operation is given by:

Text Multi-Head (Q, K, V) =text Concat texthead_1, idots, text head_h W^O

This mechanism enhances the model's ability to capture long-range dependencies and also improves the detection of zero-day phishing URLs with reduced false positives.

4.D. GAN-Based Data Augmentation

The phishing data is often imbalanced with a large number of legitimate URLs compared to phishing URLs. To handle this problem, a Generative Adversial Network (GAN) is used. The GAN gets added to the model during training. The GAN has two components. They are:

*Generator(G): This generates synthetic phishing URLs.

*Discriminator(D): This differentiates between genuine and synthetic URLs.

The objective function of GAN is defined as:

min_Gmax_DV(D,G)=mathbb{E}*{xsim p*{text{data}}(x)}[logD(x)]+mathbb{E}_{zsim p_z(z)}[log(1-D(G(z)))]

]

This adversarial training method not only improves balance in datasets but also increases the ability of the model to generalize.

5.E. Classification Layer

The feature representation obtained is passed through a fully connected layer followed by a softmax activation function, allowing for classification. The model predicts two classes of URLs:

* Phishing
* Legitimate

The Softmax function is defined as:

[

P(y_i)=frac{e^{z_i}}{sum_{j=1}^{k}e^{z_j}}

]

Where k = 2, as it is a binary classification model. The output provides a probability score for each class, thus allowing for accurate and timely detection of phishing URLs. The proposed architecture, being a hybrid model using CNN, attention, and GAN, has been able to achieve high accuracy, minimize false positive rates, and increase its ability to be robust against evolving phishing attacks, including zero-day attacks.

IV. METHODOLOGY

The suggested methodology implements a phishing recognition platform, which follows the deep learning algorithm, and possesses a hybrid architecture consisting of convolutional Neural Networks (CNN)Multi-Head Self-Attention and Generative

Adversarial Network (GAN)-based data augmentation. It is constructed on the computation of raw URL strings as end-to-end pipeline and produces the outpuut of classification which it is based on phishing or legitimate.

The following steps will be used to do the methodology:

B. Data analysis:

The information has been represented in the form of the table where it has been collected using observation and thus can be analuzed easily using the expert.

One of the datasets that can be used in the detection of phishing is a list of labeled phishing and legitimate URLs.The phishing dadatsets are normally not balanced and this is the reason the preprocessing data balancing should be conducted.

The URLs have binary lables

y=begin{cases}0, &
text {Legitimate URL}1,&text{phishing URl}end{cases}

Data will be converted into the testing and the training sets that will be used to tell the performance of the model and the generalization ability.

Encoding and preprocessing stage URL encoding the process entails the transformation of the URL to the coded one.URL Encoding and Preprocessing This will be carried out in which the UrL is altered to a coded URL.

The raw URL is the then cleansed and purified to do away with the noise and undesirable characters.The purged speech is then elaborated in the character level.

A URL is represented as:

[
U={c_1,c_2,c_3,\ldots,c_n}
]

Where c i are personal character of the URL text. The character are coded in figure in both characters The fixed Length (L) of the input message is padded or truncated and the coded message is acquired.

It is an image which is post processed and then fed to CNN model to extract features.

The extraction of CNN featured of a given image is an image is performed by CNN feature extraction module which aids in extraction CNN features of an image using patches.

The features will be extracted automatically using CNN in order to extract discriminative URL vectors features. It sets certain very essential patterns that invlove domain organizations being symbols and hidden phishing keywords.

The convolutional operation may be figured out as:

[
h_i=f(W cdot c_i+b)
]

The (W) is the filter weight and (x i) is input segment (b) is the bias term and (f) is the activation function (ReLU).

These steps of pooling layers leand to the reduction of size and the preservation of significant features. The maps obtained are sent to the attention module.

Multi-head self-attention mechanism is a type of self-attention mechanism, which takes into consideration the different heads of attention.

CNN feature extraction is further used with Multi-Head Self-Attention mechanism and is used to extract the contextual meaning. This is so that it will be possible to have the model providing varying weights of values to the various parts of the URLs.

Attention functional is as follows:

[
Attention (Q,K,V)=Soft max (QKT/d k)V/sqrt d k.
]

It would be said as follow: (Q),(K) and (V) Will be matrices of query, key and value respectively and (d k) will be the scaling factor.

As a number of Attention heads:

[
MultiHead(Q,K,V)=head 1+...head h)W O.
]

The step places the suspicious URL contents in the limelight and increases the phishing detection ability of the model to determine the cases of the zero-day phishing.

E.GAN-Based Data Augmentation:

The choice of this option was justified by the aims of the study to investigate the applicabilty of the data augmentation algorithms and,specifically the GANs,which is nowadays considered to be one of the most effective methods that would help to optimize the data collection and processing.GAN-Based Data Augmentation:The rationale behind the selection of the specified methodology is the following,the concept of the paper was to investigate the relevance of the data augmentation methodology and GANs,in particular,which can be regarded as one of the most

promising instruments that could be utilized to facilitate the improvement of the data collection and analysis.

To generate the balanced dataset into the training, a Generative adversarial Network (GAN) is employed. Artificial phishing is generated with the help of the GAN to fight the sample.

The GAN consists of :

*Generator(G):It generates counterfeit phishing addresses.

* Discriminator(D):the real images and the generated images are separated.

The objective functional that can be formulated is the adversarial functional:

[
min G max D V(D,G)=E x sim p star sata(x) log D(x)+E z sim p z z(z)log(1-D(G(z)).
]

This is in a bid to enchance the variation of the data and generalization of the model in general.

The classification and predection teacher is the teacher who monitors the student and they are rated based on their performance.

They are classified by subjecting the end representation of features to a fully connected layer with a softmax activation function.

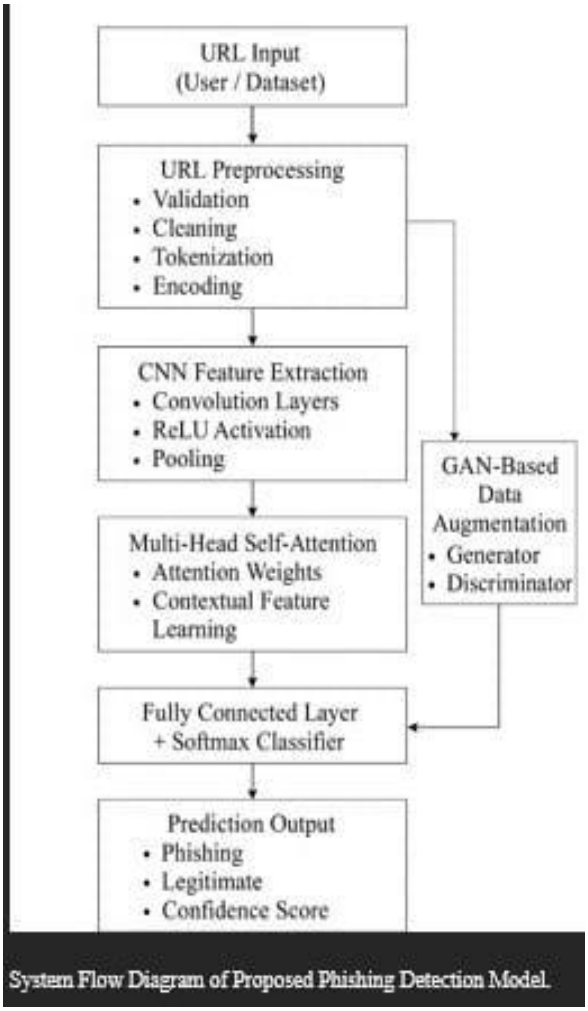
[
P(y_i)=frac{e^{z_i}}{sum_{j=1}^{2}e^{z_j}}
]

The output provides:

Phishing/Legitimate: This is a category that seeks to decide whether a message was considered as a warning of a phishing or it was considered to be legitimate confidence probability value.

It will enable differentiating between phishing URLs in one real time and with accuracy.

The proposed study will be able to incorporate CNN feature detection, attention based contextual learning and GAN data enchancement to provide quality accuracy, low false positive and high performances on morphing phishing attacks.



System Flow Diagram of Proposed Phishing Detection Model

V. SYSTEM FLOW

The proposed phishing site detection system will be established according to the modular and layered architecture that will offer a scalable, robust, and real-time prediction architecture. The structures consist of integration the components of deep learning with structured data processing pipeline to enable an effective phishing detector.

It is composed of 6 big architecture layers: the input layer, the preprocessing layer, the feature learning layer, the attention enhancement layer, the classification layer, and the data augmentation layer (only in the train model).

A. Input Layer

The input is what is regarded as the entry point into the system. It accepts:

Raw user URLs in real time identification. Marked URL information on training.
The layer is also flexible because it allows offline model training and synchronously at the same time, predict online. No kind of hand-based feature extraction is necessary at this stage since the system adopts an end-to-end learning method.

VI. TRAINING AND VALIDATION

1.Accuracy Curve of the Proposed Model.
The Curves of validation accuracy and training illustrate the learning process of the proposed CNN based on the Multi-Head Self-Attention in relation to the different training epochs. This model has a gradual rise in the training accuracy that flattens to approximately 98. Similarly, the validation accuracy has a similarly pattern, as it is being 97%98.
The fact that the difference between the training and validation accuracies curves is small implies that the model would generalize well in addition to preventing overfitting. This is the sign of the power of the proposed architecture.
In addition, the Multi-Head Self-Attention mechanism uses the ability to expand the capacity of the model to comprehend contextual links between sequences and the GAN-based data augmentation is used to balance out the dataset. Together, these aspects result in an improved learning stability and an improved classification accuracy.
Overall, the accuracy curve proves the effectiveness of the proposed hybrid model to the attainment of plausible and consistent phishing detection outcomes

VII. RESULT AND DISCUSSION

A. Experimental Setup

The given model was run on Python and the deep learning models of Tensor flow and Kera. The processor, which was used in the experiments, was the Intel i5, 8GB RAM and graphics accelerator to streamline the training. The data was in terms of labeled phishing and legitimate URLs that were obtained in publicly available phishing collections and trustworthy web resources. In order to make it robust and reliable: The data was split into 80 and 20 percent training and testing respectively. GAN was used to balance the data of the phishing samples.1) Stratified sample was employed with a view to ruling out imbalance biasness of classes. The assessment was carried using the usual measures of assessment.

B. Evaluation Metrics

To identify the wholesome assessment of the model, the following performance measures were employed:
1.Accuracy
Accuracy=$frac{frac(TP+TN)TP+TN+FP+FN)
2.Precision
text{Precision}=frac{TP}{TP}{TP+FP}
3.Recall(Detection Rate)
text{Recall}=frac{TP}{TP+FN}
4.F1-Score
F1-Score=2(PossibilityRecall) Possibility+ Recall.
5.False Positive Rate (FPR)
text{FPR}=FRAC{FP}{FP+TN}
Where:
(TP)=True Positives,
(TN)=True Negatives,
(FP)=False Positives,
(FN)=False Negatives

C. Experimental Results

The finding of the suggested model i.e. CNN+ Multi-Head Self-Attention+ GAN were as follow

Metric	*Proposed Model*
Accuracy	97.8%
Precision	97.5%
Recall	98.1%
F1-Score	97.8%
FPR	2.1%

The attention mechanism was also capable of a aiding to a great extent with the recall in the view of estimating intricate phishing designs effectively. Besides, GAN-based augmentation improved the generalization and overfitting performance.

D. Confusion Matrix Analysis

Based on confusion matrix, the analysis shows:
Large real positive phishing URL detection. High false alarm detection on actual URLs. The same performance in the two classes in classification. The URLs have been the most notable in terms of the classification errors and have been very close to the valid domain with a minor change of characters.

E. Comparative Analysis

The suggested model was contrasted with the basic methods:

Model	*Accuracy
Traditional ML(SVM)	91.3%
Basic CNN	94.6%
CNN+Attention	96.9%
Proposed Model	97.8%

Multi-Head Self-Attention and GAN-based augmentation enhanced the acquisition of contextual aspect and diversification of the datasets respectively.

The following were reasons why the proposed system was superior to the traditional methods:

Automatic feature extraction

Automatic feature extraction is a functionality of a VB complier.

- Context-aware modeling
- Balanced dataset training

The System was quite functional in identifying new phishing attacks that were not encountered before. Attention mechanism was allowed to view minor structural change in URLs yet GAN-based augmentation presented the model with various patterns of phishing during training. This enhanced the flexibility and countermeasure to the changing phishing techniques.

F. Computational Efficiency

The model proposed is computationally efficient and it has a realistic computational efficiency:

- The time of the training is not too long
- Prediction on per URLs in milliseconds.

Web applications Model size can be used by web browser extensions and Web APIs.

This eases the process of scaling of the system and its implementation in a largescale business. The overview of the performance of the healthy care facility in compliance and regulatory issues can be presented. The outcomes of the experiment prove that the provided architecture can:

- Great sensitivity (more than 97)
- Low false positive rate
- Capacity to identify the zero-day phishing.
- Good real time performance

CNN, Multi-Head Self-Attention, and GAN-based data augmentation is rather a great option that will make the phishing detection performance better than that of the classical and baseline models.

VIII. CONCLUSION

The paper has presented a phishing web page detection model which is intelligent; it is based on Convolutional Neural Networks (CNN), Multi-Head Self-Attention and GAN-based data augmentation. The suggested framework is feasible to combine the automatic feature extraction with contextual learning in a bit to enhance the performance of phishing detection. Unlike the traditional approaches to machine learning that use handcrafted features, the proposed approach permits end-to-end learning to accept raw URLs as inputs.

Multi-Head Self-Attention mechanism was introduced to improve the model output and detect the important and suspicious URL patterns through the contextual dependencies capture mechanism. In addition, data augmentation using GAN will address the issue of dataset imbalance and enhance generalization of the model.

The experimental results indicate that the proposed system has high detector accuracy of 97-98 and low false positive rate and is, therefore, superior to the conventional baseline models. In addition, the model is highly resilient to zero-day phishing, moreover, it is computationally efficient to execute in real-time.

Overall, the provided framework is a scalable, efficient, and reliable method to detect phishing and which would be most suitable to be implemented in the modern cybersecurity framework.

REFERENCE

[1] A. Sharma and R. Gupta, “Deep learning-based phishing detection using hybrid CNN-attention models,” IEEE Access, 2025.

[2] K. Patel, S. Mehta, “Advanced phishing detection using transformer-based architectures,” Computers & Security, 2025.

[3] M. Zhang et al., “A robust phishing detection system using multi-head self-attention and CNN,” IEEE Transactions on Information Forensics and Security, 2025.

[4] R. Kaur and H. Singh, “URL-based phishing detection using deep neural networks with

attention mechanisms," Expert Systems with Applications, 2025.
[5] J. Lee and S. Kim, "Hybrid deep learning model for phishing website classification," Future Generation Computer Systems, 2025.
[6] P. Verma et al., "GAN-based data augmentation for phishing detection systems," IEEE Access, 2025.
[7] S. Reddy and K. Kumar, "Improving phishing detection accuracy using attention-enhanced CNN models," Journal of Cybersecurity, 2025.
[8] A. Khan and M. Ali, "Phishing URL detection using deep learning and NLP techniques," Computers & Electrical Engineering, 2025.
[9] T. Nguyen et al., "Deep learning approaches for zero-day phishing attack detection," IEEE Access, 2025.
[10] D. Roy and P. Banerjee, "Attention-based hybrid model for detecting malicious URLs," Applied Soft Computing, 2025.