INTERNATIONAL JOURNAL OF EDUCATION, RESEARCH AND SCIENTIFIC DEVELOPMENT
https://www.ijresd.org
Print: 2992-5665 and Online: 2992-5673
Impact Factor: 5.2 Vol 4, Issue 2 December 21, 2023

A PHISHING DETECTION SYSTEM BASED ON INTELLIGENT DEEP LEARNING TECHNIQUES

Uzoaru, Godson Chetachi, [1]
[1]Department of Computer Science, Clifford University, Owerrinta, Abia State, Nigeria
uzoarugc@cliffordduni.edu.ng

Nwamuruamu Godswill, [2]
[3]Department of Computer Science, Clifford University, Owerrinta, Abia State, Nigeria
godswillnwamuruamu@gmail.com

Johnson-Okoronkwo Cynthia, [3]
[2]Sterling Bank PLC
cjohnsonokonlwo@gmail.com

ABSTRACT

In contrast to software vulnerabilities, phishing websites or URLs are a type of internet security problem that focuses on human vulnerabilities. There are many ways to compromise an internet user's security, but the most popular one is phishing, an attack that seeks to obtain or misuse a user's personal data, such as passwords, credit card information, identity, and account information. Phishers collect information about users by impersonating legitimate websites that are indistinguishable to the naked eye. Users' sensitive information may be accessed, putting them at risk of financial harm or identity theft. As a result, there is an urgent need to create a system that effectively detects phishing websites. This paper proposes three distinct deep learning-based techniques for detecting phishing websites, including long short-term memory (LSTM) and convolutional neural network (CNN) for comparison, and finally an LSTM-CNN-based approach. Experimental findings demonstrate the accuracy of the suggested techniques, i.e., 99.2%, 97.6%, and 96.8% for CNN, LSTM–CNN, and LSTM, respectively. The proposed phishing detection method demonstrated by the CNN-based system is superior.

Keywords: Intelligent, phishing detection; deep learning; convolutional neural network (CNN); LSTM; detection of cyber-attacks

1. Introduction

With the advent of more recent technological advancements, cybercrime is becoming a global problem that affects the entire world [1]. Furthermore, law enforcement organizations are working feverishly to apprehend the criminals due to the crime's widespread distribution [2]. Lawmakers are

enacting stricter legislation to gradually discourage this new form of criminal activity [3]; police departments are also training officers to be more tech-savvy [4] and developing specialized computer crime experts [5]. But politicians and law enforcement are powerless to combat the massive and pervasive problem of cybercrime [6].

As the growth of E-services expands, so do attackers' opportunities to gain or misuse users' information such as their names, phone numbers, identification, and credit card information [7]. As a result, users face a variety of online threats and cyber-attacks every day. Phishing can occur via electronic mail (E-mail), SMS (Short Message Service), or URL (Uniform Resource Locator), to name a few [8] The hacker has the ability of stealing a large amount of critical data, including account information, credit card information, user personal information, passwords, and identity

.

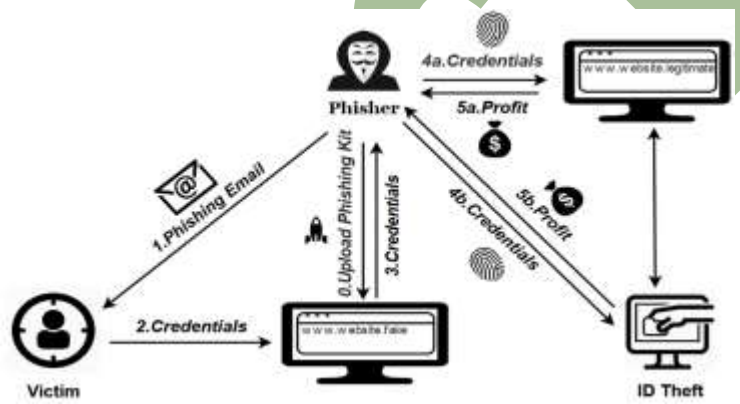

Figure 1: The illustration of a Phishing Attack: [9]

The Anti-Phishing Working Group (APWG) observed 1,286,208 phishing attacks in the second quarter of 2023. This was a decrease from the 1,624,144 attacks seen in the first quarter of 2023, which was the highest quarter in our historical observations. The APWG's 2Q 2023 total was the third-highest quarterly total ever recorded. It was significantly higher than the 888,585 in 4Q 2022 and roughly equal to the 1,270,883 in 3Q 2022. However, phishing had significantly decreased by the end of the second quarter. The 306,847 attacks recorded in June 2023 were the fewest since November 2021.

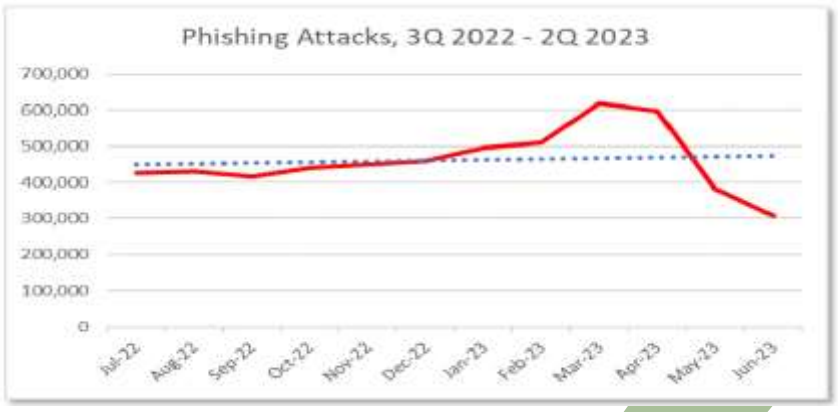

Figure 2: Phishing report for Second quarter of the year 2023 [10]

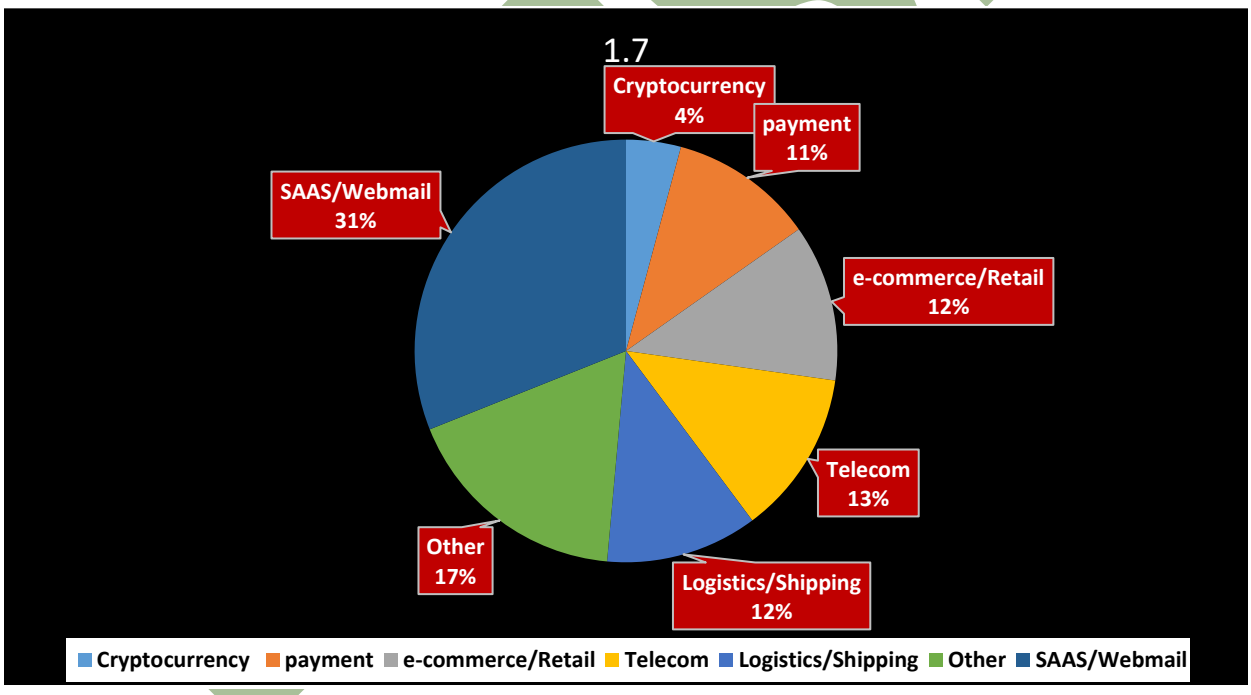

Figure 3: STATISTICS OF PHISHING ATTACK AND MOST-TARGETED INDUSTRIES, 2Q 2023

Table 1: : STATISTICS OF PHISHING ATTACK AND MOST-TARGETED INDUSTRIES, 2Q 2023

INDUSTRIES	***LEVEL OF ATTACK (%)***
Gaming	1.7
Cryptocurrency	2.2

payment	5.8
e-commerce/Retail	6.3
Telecom	6.6
Logistics/Shipping	6.1
Other	9.2
SAAS/Webmail	16.3

In the second quarter of 2023, APWG founding member OpSec Security discovered that phishing attacks against the financial sector (which includes banks) remained the largest set of attacks, accounting for 23.5 percent of all phishing – the same as in 1Q 2023 2022. Attacks against online payment services accounted for another 5.8 percent of all attacks. Attacks against social media companies have grown to become a larger share of all attacks, accounting for 22.3 percent of all attacks in Q2 2023.

Table 2: Statistical Highlights for the 2nd Quarter 2023

	APRIL	MAY	JUNE
Number of unique Phishing Web sites(attacks) detected	597,789	381572	306847
Unique phishing email campaigns	41,083	30,717	22,610
Number of brands targeted by Phishing campaigns	544	521	498

Intelligent techniques such as machine learning (ML) and deep learning (DL), which fall under artificial intelligence (AI), are rapidly evolving and effective in providing security for computing operations and cybersecurity management. AI's diverse characteristics, ranging from detecting and extrapolating patterns to providing security to adapt to a new environment, make it an essential component of technological systems such as computer vision and cybersecurity.

Human expertise is required for feature extraction and selection in classic machine learning techniques. The tasks of feature selection and classification are separated. Deep learning bridges that gap by combining both detection and classification in a single phase to optimize model performance. Deep learning models, unlike machine learning, reduce the need for manual feature engineering and reliance on third-party services due to automatic learning and feature extraction. Furthermore, the major advantages of deep learning over traditional machine learning techniques are high performance and end-to-end problem solving, particularly in cases of large datasets such as speech recognition, image classification, and phishing detection.

In different studies, ML and DL models were compared, and it was discovered that DL models performed better in terms of accuracy

for detecting phishing websites than ML models.

It is not easy to choose the best method for a given application. If the wrong algorithm or method was used, the model's accuracy and efficiency would suffer [12], especially given how frequently phishers change their methods of attack to take advantage of weak points in systems and users' ignorance. As a result, numerous anti-phishing technologies have been developed to detect phishing risks early and protect users from such attacks. Deep learning-based security mechanisms are increasingly being used in a wide range of industries to defend against emerging phishing attacks.

Applications of deep learning are employed in a variety of industries, including autonomous vehicles, facial recognition, and medical equipment. Through learning by example, deep learning trains machines to act like the brains of humans. Furthermore, using the "deep learning" process, a computer model can directly learn how to perform classification tasks from large datasets containing text, sound, and images. Deep learning models can achieve better results; in some cases, they can even outperform human performance. A large amount of labeled data, significant computing power, and neural network architectures with multiple layers are required for training deep learning models.

Because of the robustness of deep learning algorithms, researchers have proposed numerous methods for dealing with phishing websites by extracting features for classifying URLs. Numerous methods for detecting phishing attacks have been developed, including the use of different, new, and well-known features such as URL length, keyword frequency, lexical features, and the incorporation of new features.

Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that achieves superior results when dealing with time-series data by eliminating vanishing gradients and long-term dependencies.

As illustrated in Figure 4, the LSTM architecture consists of a cell and three gates (input, output, and forget).

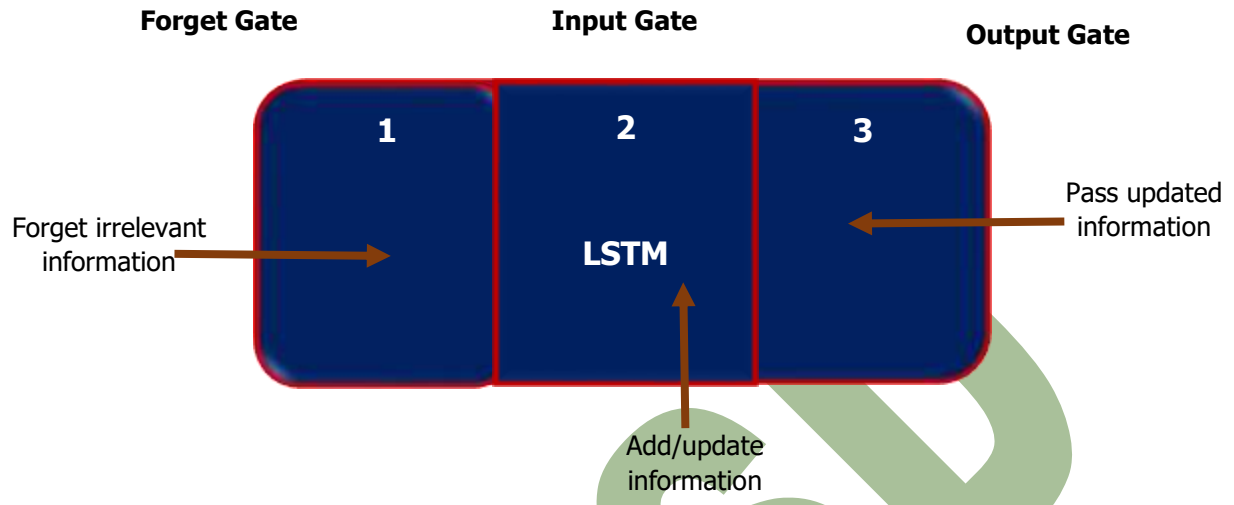

Figure 4: The fundamental architecture of LSTM [11]

A convolutional neural network (CNN) is a type of artificial neural network that is mainly used for image recognition and processing, it can also identify patterns in images. For training, millions of labelled data points are needed. CNN is useful for diagnosing illnesses and for phishing detection, image classification, and object recognition.

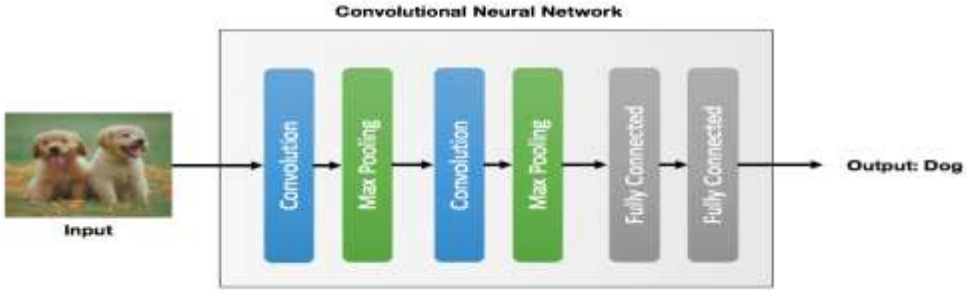

Figure 5: The fundamental architecture of CNN.

The convolution and pooling layers that make up the first few layers of a typical CNN are depicted in the following diagram: Fully Connected layers, also known as dense layers, with sigmoid or softmax activation functions are always the last layers in a CNN.

It should be noted that multiclass classification problems use the softmax activation function, while binary classification problems use the sigmoid activation function. [12]

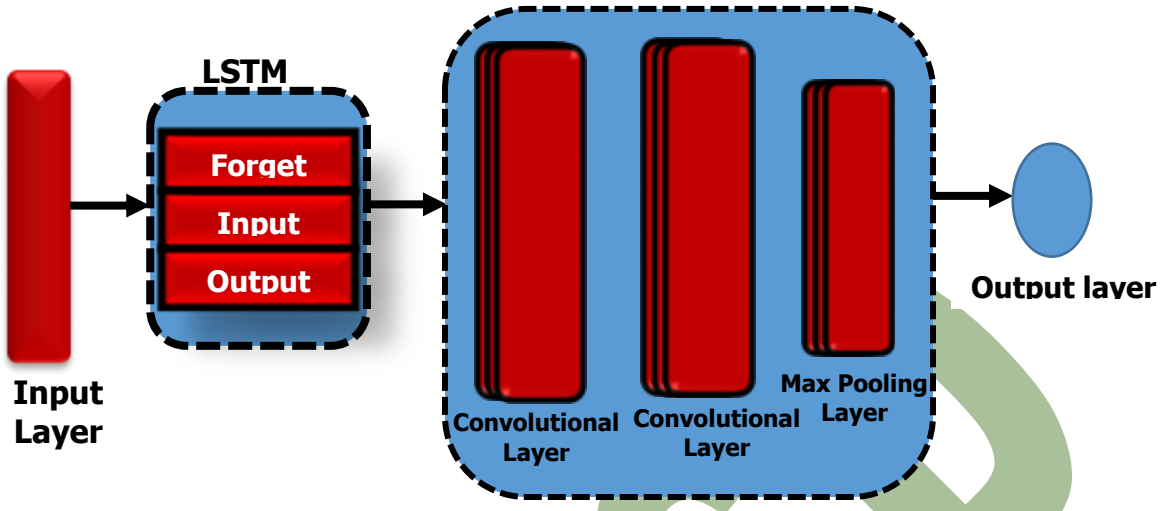

Figure 6: the LSTM-CNN Architecture

As demonstrated in Figure 6, the LSTM–CNN architecture combines the advantages of both CNN and LSTM techniques to achieve superior performance. Using CNN and LSTM for the phishing detection task is promising because they demonstrate high performance in overcoming tasks related to classification, detection, and recognition.

We were therefore inspired to develop an effective deep learning solution for phishing websites. To achieve excellent results in phishing detection, this paper employed an empirical method to examine the performance of the three techniques: LSTM, CNN, and LSTM–CNN. The objective of this research was to classify whether the URL was phished or legitimate by using LSTM, CNN, and LSTM–CNN. To ascertain whether the URLs are phished or legitimate, we propose a phishing detection system based on deep learning techniques.

2. Literature Review

The phishing website problem is complex and difficult to solve because no definitive solution exists to effectively eliminate all threats. Deep learning-based phishing website detection solutions have emerged to detect phishing websites. Furthermore, deep learning has demonstrated great potential in tackling threats in cyber security. Table 1 shows several previous works that use deep learning approaches for phishing website detection.

2.1. Long Short-Term Memory (LSTM)

Yang et al. proposed a new method for detecting phishing attacks that employs the LSTM deep learning method and optimizes model training with the combined characteristics of RNN. The main benefits of using LSTM are its ability to incorporate large amounts of data and learn complex

features automatically. This solves a complex problem that other machine learning methods are unable to solve. Yahoo and PhishTank datasets were used. [13]

2.2. Convolutional Neural Network (CNN)

A deep learning model proposed in [14] used a character-level CNN to detect phishing URLs. The study used CNN at the character level to learn the URL's sequential information, then max-pooling was used to determine important features, which were then fed to fully connected layers for classification. The stochastic gradient descent algorithm (SGD) was used to train the network. According to the results, the proposed model achieved an accuracy of 95.02% on the given dataset. Furthermore, the model's accuracy on benchmark datasets was 98.58%, 95.46%, and 95.22%, outperforming current phishing URL models and various machine and deep learning algorithms. [15] A deep learning-based phishing detection system was presented for preventing phishing attacks. There were 37,175 phishing URLs and 36,400 legitimate URLs in the dataset. CNN was used to conduct the research. The advantage of this system is that no feature engineering is required because the CNN automatically extracts features from URLs via its hidden layers. The framework consists of the input text being passed through the embedding layer, followed by the creation of a matrix and its transmission to CNN.

The proposed system achieved 98.00% accuracy. It was possible to identify a two-dimensional code phishing attempt by using the relative detection method that was proposed. The FlickrLogos-32 dataset, a publicly available logo dataset containing 32 distinct logo brands, was the source of information. In order to conduct the study, the conventional method—an enhanced feature pyramid network (FPN) coupled with a quicker R-CNN logo identification technique was improved. The three logo processes—identification, recognition, and extraction were the system's primary functions. The process of removing logos from two-dimensional code is called logo extraction [16]. Faster R-CNN was used for the logo identification and recognition based on the retrieved logos. Evaluating the logo's consistency between the identified object's described identity and the actual identification is the last step in the identification process. The results showed the method's efficacy in logo recognition when compared to other phishing detection and logo recognition methods. This method can be applied to two-dimensional code phishing assault detection.

A deep learning platform called HTMLPhish uses data-driven, end-to-end automatic classification of phishing web pages. Over 50,000 HTML documents make up the dataset, and the entire collection of HTML contents was distributed in the real world. Using a web crawler, the data were extracted from HTML documents. In order to learn the pertinent feature representations, HTMLPhish used CNNs to discover semantic dependencies in the textual content of HTML documents. Furthermore, convolutions on a combination of the word and character embedding matrix were applied

to guarantee that new words were successfully added to the test HTML documents. This method could analyze context features from HTML pages without considering laborious manual feature engineering. The results showed that HTMLPhish performed satisfactorily, with an accuracy rate of more than 93%. [17]

The vulnerability of internet users to security vulnerabilities and cyberthreats led to the development of artificial intelligence-based algorithms using machine learning and deep learning techniques. It was possible to build a system that uses a CNN with n-gram features to identify phishing and prevent cyberattacks. The system determines which n-gram feature extraction technique is more effective and which parameter works best by extracting these features from URLs. Single characters yield the greatest results. 34s is obtained for training one epoch and 0.008 s for URL classification when 70 characters are used in the model training process. It is very good to achieve an accuracy of about 88.90% with the high-risk URL dataset. [18]

A novel deep learning architecture called Texception [19] can determine if an input URL is a phishing link or not. Texception differs from classical methods in that it relies less on manually created features and instead uses character-level and word-level information from the URL. Texception expands through various parallel convolutional layers, becoming deeper or wider. Texception generalizes better for new URLs that use the Microsoft SmartScreen service dataset. Production data results demonstrated Texception's outstanding performance. With a false-positive rate of (0.01%), the true positive rate rose by 126.7%.

Because phishing websites are growing faster than ever, there is a greater need to strengthen cyber defense and implement effective phishing detection in order to deal with the increased exposure to various cyberattacks. A 1D CNN-based model was employed, which makes use of CNN's ability to distinguish between phishing and genuine websites. The model assessed a dataset of websites that included 4898 phishing and 6157 genuine websites, according to the authors. The model makes use of unseen phishing websites detection. Additionally, the model's F1-score increased to 0.976 and its phishing detection rate to 98.2%, respectively. [20]

2.3 LSTM and CNN integration

The focus of the analysis was on how well various deep learning algorithms identified phishing websites, in order to help organizations select and implement appropriate solutions that meet their technological requirements. 11,055 benign and phishing URLs can be found in the data. They used DNN, CNN, LSTM, gated recurrent unit (GRU), and other deep learning algorithms. The model was tested on various architectures for each deep learning algorithm to determine the ideal parameter to obtain good accuracy. The outcomes showed that the optimal measure of overall performance metrics is obtained by a deep learning algorithm. [21]

Deep learning techniques can help with both natural language and image classification. In order to investigate the possibility of

differentiating phishing URLs from distinct authentic URLs, an intelligent phishing detection system (IPDS) was suggested. LSTM and CNN are used by IPDS to create a hybrid classification model. Approximately one million authentic and fraudulent URLs were employed in the dataset gathered from PhishTank and Common Crawl. Over 10,000 images and one million URLs were used in training for the CNN classifier and LSTM to create the IPDS. The number of misclassifications, split issues, and feature type were among the factors that influenced IPDS sensitivity. 93.28% classification accuracy was attained by IPDS. [22]

Many phishing detection techniques have detection rules that are computationally expensive and challenging to update in response to shifts in attack trends. PhishTrim is a deep representation learning-based, lightweight technique for detecting phishing URLs that was presented by Zhang et al. The initial embedding representation of the URLs was obtained using the skip-gram pretraining model. Moreover, Bi-LSTM was employed to extract context dependency and discover the deep representation of URLs. CNN was utilized to extract the local n-gram features using the PhishTrim dataset [23].

Because more people are shopping and banking online, hackers are able to obtain sensitive data about their victims by impersonating reputable websites and using various techniques to obtain personal information. An anti-phishing system built on CNN and LSTM was proposed in order to shield users from situations like these. The dataset included almost 200,000 URLs that were retrieved via the Yandex search API, VirusTotal, and PhishTank. The suggested system operates effectively, achieving 97% accuracy and 97% precision. Because the model has a straightforward user interface, it can be used in web browsers. [24]

Following an extensive review of the literature, phishers are constantly coming up with effective ways to get around the detectors that are in place, making phishing detection research a difficult undertaking. Studies on phishing detection techniques can be divided into groups based on the inputs they use, including HTML content, URLs, emails, screenshots, and logos. When it comes to the URL as input, the majority of studies have demonstrated that characteristics like URL length, character count, keyword frequency, and frequency of auspicious symbols all have positive correlations with datasets gathered from open phishing platforms like PhishTank, OpenPish, and VirusTotal. These investigations' findings demonstrated that deep learning-based techniques, primarily DNN, CNN, and LSTM, could achieve accuracy of 90% or higher.

Therefore, a system that can efficiently and effectively assist in detecting phishing URLs is required. Recently, deep learning has gained more attention because of its effectiveness and capacity to automatically learn features without the need for human feature engineering. In order to live up to those claims, we employed deep learning techniques to identify phishing URLs using LSTM, CNN, and LSTM–CNN, demonstrating how well they performed in this regard. To the best of our knowledge, no prior study has compared the outcomes of the

three DL methods. There are 20,000 URLs in the dataset used in this work, 9800 of which are phishing URLs [25]. We extracted the most discriminative features for the dataset and suggested using a lightweight CNN-based model for the accurate detection of phishing websites, which proved to be beneficial to the improvement of phishing detection performance. This is the main distinction between our approach and the previously mentioned deep learning-based ones.

3. Methodology

Finding phishing URLs is a crucial cybersecurity task. Because attackers craft sophisticated URLs, phishing URLs frequently appear to users as authentic URLs. Attackers may then utilize this access to obtain users' personal information for their own purposes. In this paper, a method for identifying phishing in URLs is proposed. Two distinct approaches were used in the system's implementation to find instances of phishing in URLs. A detailed explanation of the deep learning approach, dataset, training and testing procedures, and methodology is provided in the section that follows.

3.1 Proposed System

The key components of the models' configuration are covered in this section. The model's framework consists of four stages, as illustrated in Figure 7. The first stage deals with the features of the URLs, which are acquired from the dataset; the second stage is pre-processing, whereby we used SelectKBest to detect null values and scaling values of feature selection, which contributes most to the target variable; the third stage is the training of three distinct models, namely LSTM, CNN, and LSTM–CNN, by building a deep learning approach; and finally, the fourth stage is the classification of the webpage URLs as legitimate or phishing.

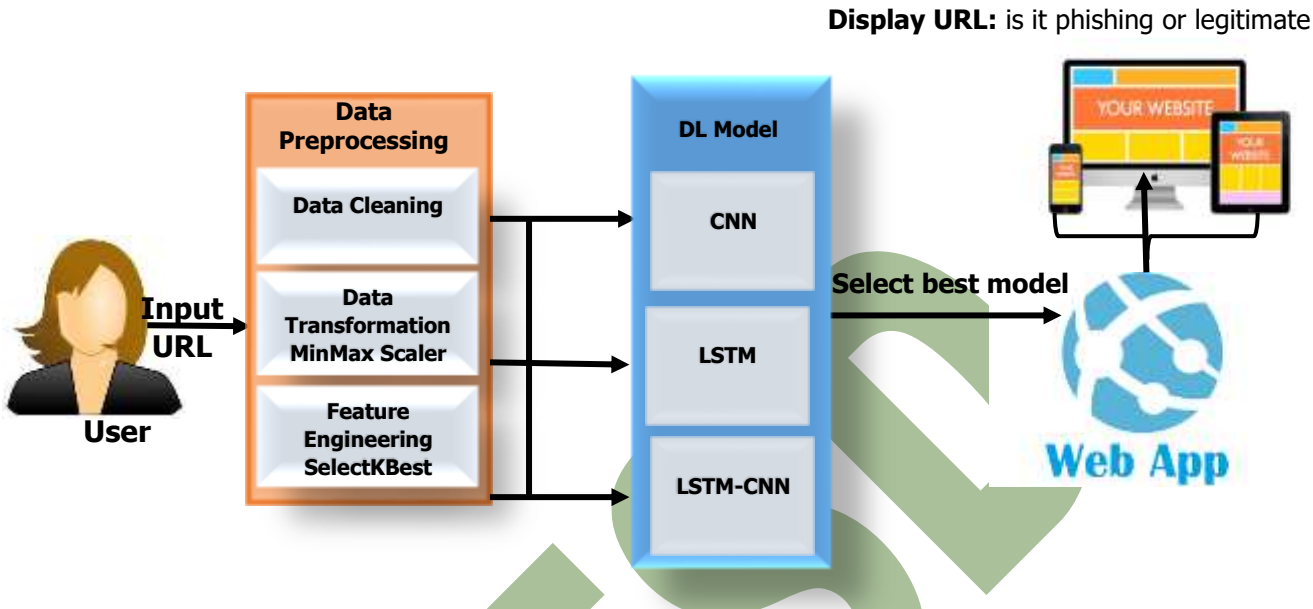

Figure 7: the Framework of the proposed system

3.2. Dataset Preparation and Preprocessing

Research validity and reliability are greatly impacted by data collection, which is why our approach's use of appropriate and consistent data ensures that the system's training is robust. After prepossessing the dataset containing the URL features, with 20,000 records of 80 features, there were many features in the dataset; thus, the value of the 30 best features was used in the SelectKBest method. The dataset under consideration was processed in the data preprocessing stage, which included scaling each feature to a given range using the MinMaxScaler method and detecting null values.

3.3. Training and Testing

The dataset was split up with 80% designated for training and 20% for testing. Table 2 displays the distribution of training and testing sets. The choice of hyperparameters made during training is one factor influencing the efficiency of deep learning algorithms. The accuracy of phishing website detection models can be increased by optimizing the values of hyperparameters. These parameters include the quantity of neurons in each layer, the number of layers overall, the batch size, the learning rate, the dropout rate, the number of epochs, the type of optimizer, the type of activation function, and the learning and dropout rates [26].

Table 3: Chart showing dataset distribution for Training and testing

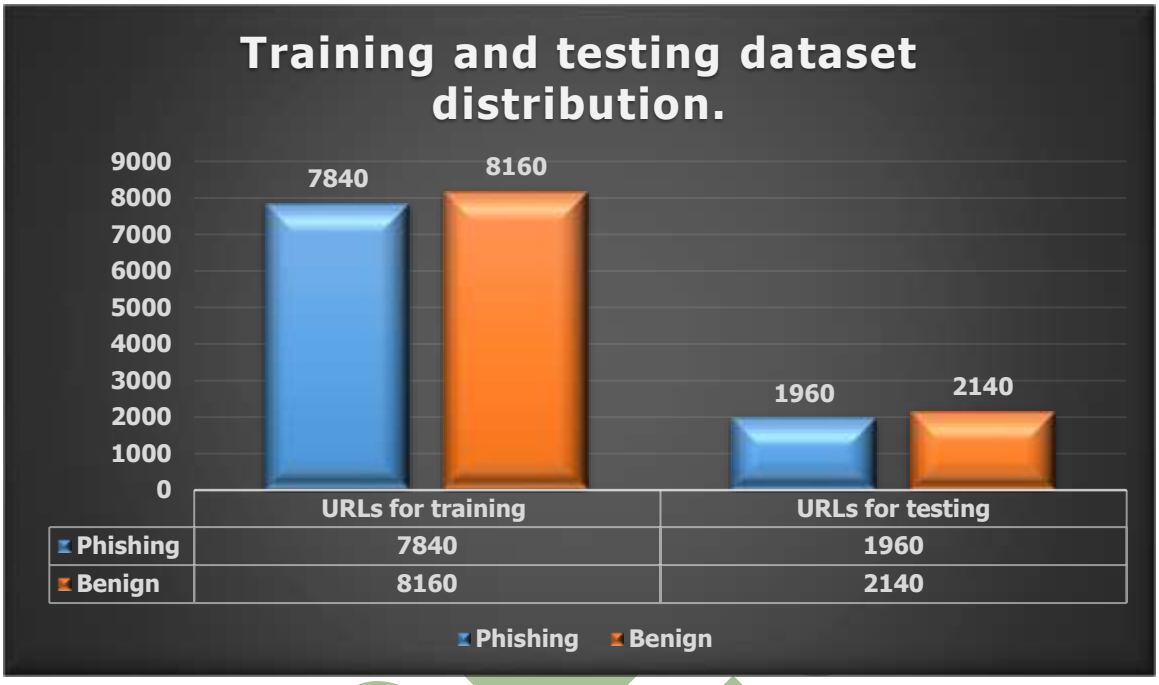

Table 4: Parameters.

	Parameters	Values
1.	Activation Function	Relu
2.	Epochs	50
3.	Batch size	1200
4.	Optimizer	Adam
5.	Dropout	0.2

3.4 Deep Learning Approaches

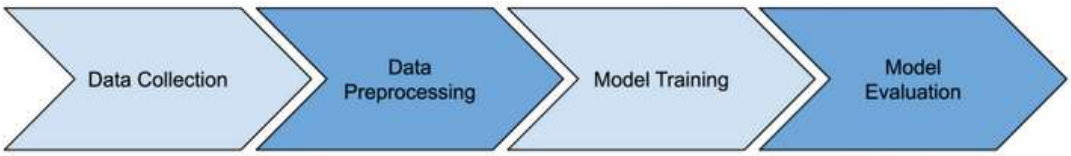

Figure 8: DL process. [27]

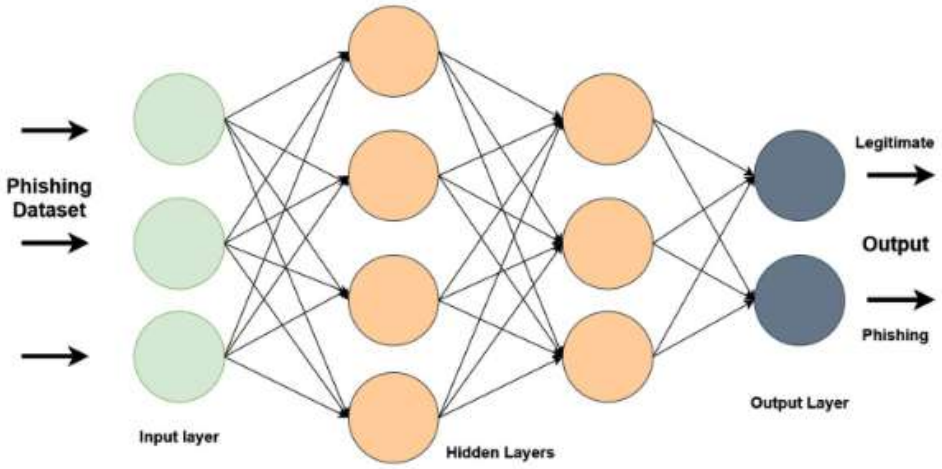

Figure 9: Using deep learning to detect phishing attacks

According to recent developments in deep learning techniques, deep neural networks (NNs) should perform better than conventional machine learning (ML) algorithms in the classification of phishing websites. However, the choice of various learning parameters has a significant impact on the outcomes of using deep neural networks [28]. Various deep learning (DL) techniques are employed in cybersecurity intrusion detection [29]. These include deep neural networks, feed-forward deep neural networks, recurrent neural networks, convolutional neural networks, restricted Boltzmann machines, deep belief networks, deep auto-encoders, and (3) recurrent neural networks. The operation of deep learning models is depicted in Figure 5. The neurons are fed a batch of input data and given weights to help them predict whether the traffic is legitimate or phishing-related.

The authors of Benavides et al. [30] attempt to combine the classification with each selected work. They describe the DL computations selected for each configuration, and the results show that the Deep Neural Network (DNN) and Convolutional Neural Network (CNN) are the most often used models out of all of them. Many DL techniques have been discussed and examined, but there is still a lack of research on the application of DL computations to the detection of cyberattacks.

The authors of Shie [31] examined various methods and discussed various approaches for accurately identifying phishing attacks. Among the assessed approaches, feature extraction-based DL procedures perform well due to their high accuracy and robustness. Classification models show good performance as well.

In order to ensure security at a vertical scale rather than even execution, authors in Maurya and Jain [32] presented an anti-phishing structure that relies on using a

phishing identification model dependent on DL at the ISP's level. This methodology is set between different workers and end-clients and includes a transitional security layer at ISPs. The effectiveness of this structure's implementation resides in the way that a single blocking goal can ensure that a sizable client base is shielded from a particular phishing attack.

The computational overhead associated with phishing discovery models is limited exclusively to ISPs, and end users receive secure support regardless of their framework designs in the absence of extremely powerful processing equipment.

To detect phishing websites, various classification methods are used, which are then evaluated using various performance metrics. In this study, three models were investigated: LSTM, CNN, and LSTM-CNN. Convolutional layers are distinguished by their ability to learn internal representations and retrieve meaningful data knowledge, whereas LSTM networks are effective at detecting both short- and long-term dependencies. According to the experimental results, the CNN model performs exceptionally well. Furthermore, each of the three models is explained below.

LSTM (long short-term memory):

Long short-term memory is an adaptive recurrent neural network (RNN), a type of recurrent neural network in which, in addition to the conservative neuron, a memory cell switches each neuron based on an internal state. The layers of LSTM are made up of memory blocks that are linked together repeatedly; each block contains one or more memory cells with recurrent connections.

As a result, a typical LSTM cell has an input gate that controls data input from outside the cell and determines whether the data in the internal state is kept or ignored, as well as an output gate that prevents or enables the ability to view the inner state from the outside [33].

Convolutional neural network (CNN):

CNN is an efficient discriminative architecture for processing two-dimensional grid-based data, such as pictures and videos. The CNN performs better than the neural network (NN) in terms of time delay. In the CNN, the weights are shared in a temporal dimension, which shortens computation time. Thus, the CNN replaces the generic matrix multiplication found in the standard NN. Consequently, the CNN method reduces the weights, which in turn reduces the complexity of the network.

The first stage in the CNN's process for classifying a URL is to retrieve the URLs' labeled training data. Next, it randomly splits the data into train and test sets. The data was finally trained by building the CNN's architecture, which included the input, output, and layers, after the training and test data had been prepared. Following every convolution, a max-pool layer was added in order to extract the key components and turn them into feature vectors. Dropout regularization was then incorporated to make sure the model did not overfit. When this

layer employs a sigmoid function, the model categorizes the output that is produced.

LSTM—CNN:

The model is made up of LSTM layers that forecast sequences and CNN layers that extract features from input data [34]. Furthermore, compared to models that only use LSTM layers, a study [35] found that combining an LSTM layer and a 1D convolution layer improves the accuracy of malicious URL identification. As a result, we decided to train the URL features of the system using 1D CNN and LSTM architecture.

Figure 7 illustrates the CNN-LSTM workflow: the dataset is preprocessed, then split into train and test sets. Data normalization is then applied before the data is fed into the model. Finally, the model is passed to the CNN and LSTM layers, along with the dense layer to prevent the dataset from being overfit. Finally, the model classifies the output produced by this layer when a sigmoid function is applied.

4. Evaluation and Results

The Proposed system is assessed in this section, and the findings are shown.

4.1. Measures of Assessment

The metrics used to assess the effectiveness of the deep learning techniques are compiled in this section. Typically, machine learning prediction algorithms are assessed based on the classification algorithm's output. Metrics such as recall confusion matrix, precision, and system accuracy were used in this study to assess the prediction results and estimate the system [36].

Precision: The number of phishing websites that are accurately identified as genuine phishing websites represents the prediction algorithm's precision.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} = \frac{\boldsymbol{True\ positive}}{\boldsymbol{Total\ Predicted\ Positive}} \quad \textbf{(1)}$$

Recall: The number of accurate phishing URL predictions made across all URLs in the dataset is the prediction algorithm's recall.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} = \frac{\boldsymbol{True\ positive}}{\boldsymbol{Total\ Predicted\ Positive}} \quad \textbf{(2)}$$

Accuracy: The accuracy of the prediction algorithm is the ratio of the total number of correct predictions of class to the actual class of the dataset. Equation (3) calculates the accuracy of the model. Typically, any prediction model produces four different

results, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$Acuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

F1-Score: the method of calculating the precision and recall of a classifier using the harmonic mean. It is able to be merged into one metric.

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

4.2. Results

We computed the prediction algorithms' accuracy, precision, recall, and F1 score based on the experimental results. Prediction model accuracy, which has been identified as one of the common performance measures, was used to evaluate the proposed system in the majority of prediction models. Section 3 contains the prediction accuracy of the methods used in this paper. We made use of a dataset with 20,000 URL records, each with 80 features. We identified null values and scaled features during the preprocessing phase. Using SelectKBest, we then selected 30 features, on the basis of which we trained the CNN, LSTM, and LSTM–CNN classifiers.

Good results from the three suggested methods are displayed in Table 5, which also shows the best parameter selection. Following the implementation, training, and testing of the LSTM, CNN, and LSTM–CNN techniques, the results indicated a degree of improvement in phishing detection through the use of the CNN algorithm, which achieved the highest accuracy at 99.2%. The LSTM–CNN algorithm came in second with 97.6% prediction accuracy, while LSTM achieved 96.8%, as shown in Figure 10. CNN is better than the other two models for various reasons, including the fact that it performs better in terms of accuracy and other performance metrics: First, LSTM performs better for sequential data because it can learn texts and the relationships between tokens more effectively than CNN can, especially when it comes to text classification problems. Additionally, CNN is faster and more efficient than the LSTM-based method.

In addition, the model's complexity is decreased because it requires fewer training parameters than LSTM. Moreover, CNN operates at a speed one order of magnitude quicker than LSTM and LSTM–CNN. Lastly, while LSTM captures the dependency across time sequences in the input vector, CNN computations can happen in parallel.

Table 5: The outcome of the performance.

Evaluation Metric	LSTM (%)	CNN (%)	LSTM-CNN (%)
Accuracy	96.8	99.2	97.6
Precision	95.9	99	96.9
Recall	97.5	99.2	98.2
F1-score	96.8	99.2	97.6

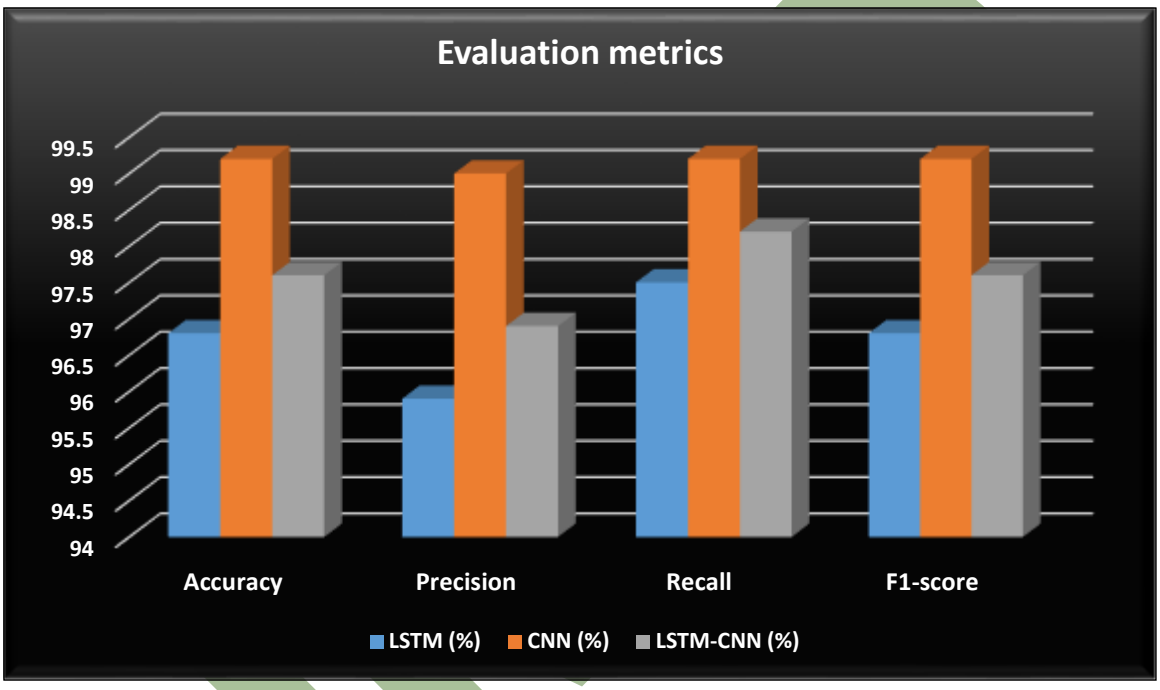

Figure 10: Evaluation metrics

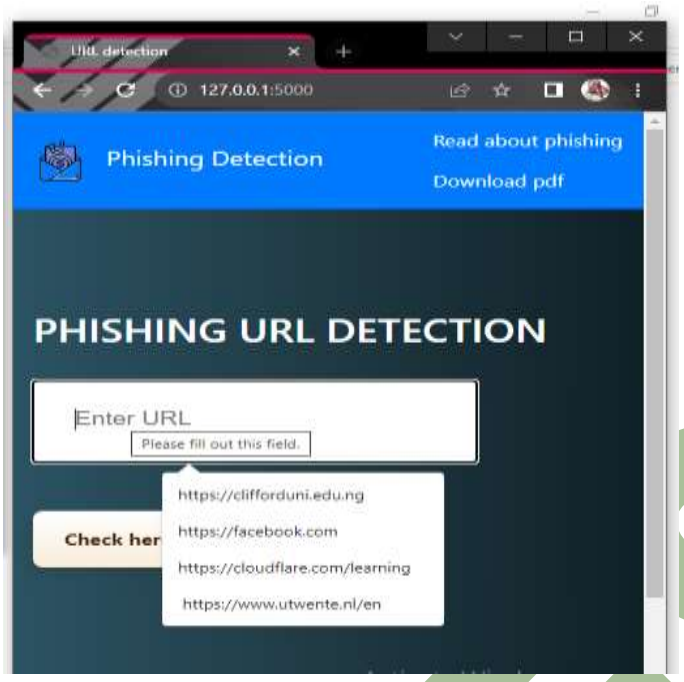

Figure 11: Sample of input Intelligent Phishing detection software

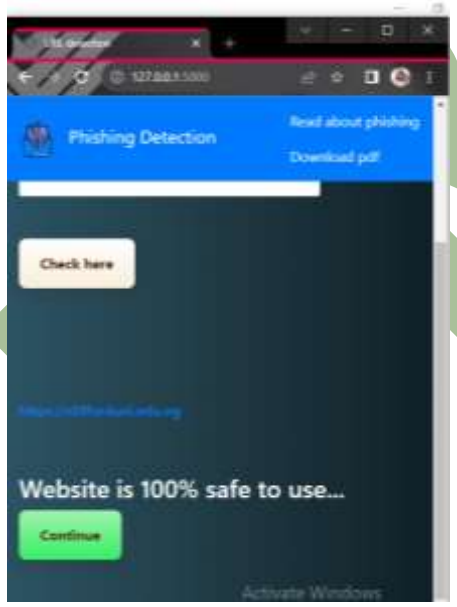

Figure 12: Sample of output Intelligent Phishing detection software

Limitations

We can see that our suggested system outperforms current approaches and

produced excellent results after testing and evaluation. The suggested system, however, is not without flaws. The model's failure to verify the website's URL status, or whether it is active or not, affects the outcomes. In order to get around this restriction, we might need to enhance feature engineering and expedite the training process. This would enable us to confirm the current status of the website and raise the accuracy of the training process.

6. Conclusions and Future Work

The proliferation of online transactions and purchases has been greatly aided by technological advancements, which simplify our daily lives. However, when sensitive information is exchanged online, it can result in unauthorized access to the data of individuals, businesses, or users. The most crucial component in defending users against information theft by phishers during online communication is security. One known method of obtaining user information is phishing, which uses a URL that appears exactly like the webpage in question. One important step in preventing hackers from accessing user data is identifying phishing attacks. Given that the number of victims is increasing due primarily to ineffective adoption of security technologies, users must be protected from cyberattacks with a clever strategy. Deep learning has demonstrated to be a valuable advancement in comparison to conventional signature-based and classic machine learning-based solutions because of its high performance and end-to-end problem-solving capabilities, despite the rapid development of deep learning techniques.

The LSTM, CNN, and LSTM–CNN algorithms were presented in this work to identify and categorize website URLs as either authentic or phishing. The evaluation of the suggested system showed that phishing website detection produced excellent results. The performance of the suggested deep learning algorithms on the same dataset varied. In terms of accuracy, the CNN algorithm performed better than LSTM–CNN and LSTM, reaching 99.2%, compared to 97.6% and 96.8% for LSTM–CNN and LSTM, respectively. In order to verify the states of websites and increase the overall accuracy of training processes, our future goals include shortening training times and optimizing feature engineering.

Additionally, we plan to present a method for identifying phishing websites that takes into account both the URL and the context of the webpage.

REFERENCES

[1]. Sun, Y., Xue, Z., Liu, Q., Jia, Y., Li, Y., Liu, K., & Su, C. Y. (2021). Modulating electronic structure of metal-organic frameworks by introducing atomically dispersed Ru for efficient hydrogen evolution. *Nature communications*, *12*(1), 1369.

[2]. Cremer, F., Sheehan, B., Fortmann, M., Kia, A. N., Mullins, M., Murphy, F., & Materne, S. (2022). Cyber risk and cybersecurity: a systematic review of data availability. *The Geneva Papers on risk and insurance-Issues and practice*, *47*(3), 698-736.

[3]. Estay, D. A. S., Sahay, R., Barfod, M. B., & Jensen, C. D. (2020). A systematic review of cyber-resilience assessment frameworks. *Computers & security*, *97*, 101996.

[4]. Parvin, K., Hannan, M. A., Mun, L. H., Lipu, M. H., Abdolrasol, M. G., Ker, P. J., & Dong, Z. Y. (2022). The future energy internet for utility energy service and demand-side management in smart grid: Current practices, challenges and future directions. *Sustainable Energy Technologies and Assessments*, *53*, 102648.

[5]. Liu, M., Yeoh, W., Jiang, F., & Choo, K. K. R. (2022). Blockchain for Cybersecurity: systematic literature review and classification. *Journal of Computer Information Systems*, *62*(6), 1182-1198.

[6]. Swessi, D., & Idoudi, H. (2022). A survey on internet-of-things security: threats and emerging countermeasures. *Wireless Personal Communications*, *124*(2), 1557-1592.

[7]. Chaimaa, B., Najib, E., & Rachid, H. (2021). E-banking overview: concepts, challenges and solutions. *Wireless Personal Communications*, *117*, 1059-1078.

[8]. Azeez, N. A., Salaudeen, B. B., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2020). Identifying phishing attacks in communication networks using URL consistency features. *International Journal of Electronic Security and Digital Forensics*, *12*(2), 200-213.

[9]. Forecast. (2017). Global fraud and cybercrime forecast. https://www.rsa.com/en-us/blog/2016-12/2017-global-fraud-cybercrime-forecast. Accessed from 20 July 2020

[10]. (2023). Apwg trend report. http://docs.apwg.org/reports/apwg_trends_ report_q4_2016. pdf. Accessed from 26 November 2023

[11]. 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.

[12]. Goyal, M., Goyal, R., Venkatappa Reddy, P., & Lall, B. (2020). Activation functions. *Deep learning: Algorithms and applications*, 1-30.

[13]. Su, Y. Research on Website Phishing Detection Based on LSTM RNN. In Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 12–14 June 2020; pp. 284–288.

[14]. Tajaddodianfar, F., Stokes, J. W., & Gururajan, A. (2020, May). Texception: a character/word-level deep learning model for phishing URL detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2857-2861). IEEE.

[15]. Singh, S., Singh, M. P., & Pandey, R. (2020, October). Phishing detection from URLs using deep learning approach. In *2020 5th international conference on computing, communication and security (ICCCS)* (pp. 1-4). IEEE.

[16]. Abdulrahman, L. M., Ahmed, S. H., Rashid, Z. N., Jghef, Y. S., Ghazi, T. M., & Jader, U. H. (2023). Web Phishing Detection Using Web Crawling, Cloud Infrastructure and Deep Learning Framework. *Journal of Applied Science and Technology Trends*, *4*(01), 54-71.

[17]. Abdulrahman, L. M., Ahmed, S. H., Rashid, Z. N., Jghef, Y. S., Ghazi, T. M., & Jader, U. H. (2023). Web Phishing Detection Using Web Crawling, Cloud Infrastructure and Deep Learning Framework. *Journal of Applied Science and Technology Trends*, *4*(01), 54-71.

[18]. Korkmaz, M.; Kocyigit, E.; Sahingoz, O.K.; Diri, B. Phishing Web Page Detection Using N-gram Features Extracted From URLs. In Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 11–13 June 2021; pp. 1–6.

[19]. Asiri, S., Xiao, Y., Alzahrani, S., Li, S., & Li, T. (2023). A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks. *IEEE Access*.

[20]. Yerima, S.Y.; Alzaylaee, M.K. High Accuracy Phishing Detection Based on Convolutional Neural Networks. In Proceedings of the 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, Saudi Arabia, 19–21 March 2020; pp. 1–6.

[21]. Jafar, M. T., Al-Fawa'reh, M., Barhoush, M., & Alshira'H, M. H. (2022). Enhanced Analysis Approach to Detect Phishing Attacks During COVID-19 Crisis. *Cybernetics and Information Technologies*, *22*(1), 60-76.

[22]. Adebowale, M.; Lwin, K.; Hossain, M. Intelligent phishing detection scheme using deep learning algorithms. J. Enterp. Inf. Manag. 2020

[23]. Zhang, L.; Zhang, P. PhishTrim: Fast and adaptive phishing detection based on deep representation learning. In Proceedings of the 2020 IEEE International Conference on Web Services (ICWS), Beijing, China, 19–23 October 2020; pp. 176–180.

[24]. Janet, B.; Reddy, S. Anti-phishing System using LSTM and CNN. In Proceedings of the 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangaluru, India, 6–8 November 2020; pp. 1–5.

[25]. 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.

[26]. Mahdavifar, S.; Ghorbani, A. Application of deep learning to cybersecurity: A survey. Neurocomputing 2019, 347, 149–176

[27]. S. Asiri, Y. Xiao, S. Alzahrani, S. Li and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," in *IEEE Access*, vol. 11, pp. 6421-6443, 2023, doi: 10.1109/ACCESS.2023.3237798.

[28]. Vrbanˇciˇc, G., Fister Jr, I., & Podgorelec, V. (2018). Swarm intelligence approaches for parameter setting of deep learning neural network: Case study on phishing websites classification. In Proceedings of the 8th international conference on web intelligence, mining and semantics (pp. 1–8)

[29]. Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419.

[30]. enavides, E., Fuertes, W., Sanchez, S., & Sanchez, M. (2020). Classification of phishing attack solutions by employing deep learning techniques: A systematic literature review. In Developments and advances in defense and security (pp. 51–64). Springer.

[31]. Shie, E. W. S. (2020). Critical analysis of current research aimed at improving detection of phishing attacks. Selected computing research papers, p. 45.

[32]. Maurya, S., & Jain, A. (2020). Deep learning to combat phishing. Journal of Statistics and Management Systems, pp. 1–13.

[33]. Adebowale, M.A.; Lwin, K.T.; Hossain, M.A. Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection. In Proceedings of the 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Island of Ulkulhas, Maldives, 26–28 August 2019; pp. 1–8.

[34]. Ariyadasa, S.; Fernando, S.; Fernando, S. Detecting phishing attacks using a combined model of LSTM and CNN. Int. J. Adv. Appl.Sci. 2020, 7, 56–67.

[35]. Pham, T.; Hoang, V.; Ha, T. Exploring Efficiency of Character-level Convolution Neuron Network and Long Short Term Memory on Malicious URL Detection. In Proceedings of the 2018 VII International Conference on Network, Communication and Computing–ICNCC 2018, Taipei City, Taiwan, 14–16 December 2018.

[36]. Lakshmi, V.; Vijaya, M. Efficient prediction of phishing websites using supervised learning algorithms. Procedia Eng. 2012, 30, 798–805.