**An Efficient Phishing Detection System based on Deep Learning Approaches**

Abstract

In terms of the internet and communication, security is the fundamental challenging aspect. There are massive ways to injure the security of worldwide internet users. Phishing is a form of attack which is focused on stealing or misusing the users' personal information such as account details, personal information, passwords, and credit card information. Previous studies sought to solve this problem; however, many studies still have obstacles which include, the use of small datasets which affect the accuracy of the proposed systems. Therefore, there is a need for a system that shows its robustness with a large dataset containing huge records and variations in features. Phishing detection methods are needed to ensure the security of internet users. Deep learning shines when it comes to complex problems, big datasets, a lot of features, and solving the problem end to end. Deep learning is getting popularity and supremacy in terms of accuracy due to achieving high results as state-of-the-art accuracy. In this work, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) algorithms-based classification, and an LSTM-CNN are proposed to detect phishing websites. A dataset with 15368 records and 80 features is used. Experimental results show the accuracy proposed system as 94%, 92%, and 91% for LSTM-CNN, CNN, and LSTM respectively.

**Keywords:** phishing detection; website URL; deep learning; cyber-attack detection.

# Introduction

Life becomes more accessible and fast because of the evolution of E-services, digitalization, and electronic communications, especially during the lockdown because of the Covid-19 pandemic when all transactions and life needs must hold online than doing it physically. To do your daily needs you can simply open your smart device, and search for the website you want such as a pharmacy, shop, learning platform, or bookstore. On the other hand, the growth of E-services expands the attacks' opportunities to gain or misuse users’ information such as their names, phone numbers, ID, and credit card information. As a result, Users face a variety of online threats, and cyber-attacks every day. Phishing has different types, it could be done via E-mail, SMS (Short Message Service), or URL (Uniform Resource Locator) to name a few. Phishing can compromise all types of data sources including personal information, and online accounts, and gain access and modification to connected systems.

In some cases, hackers stop phishing when they steal enough information for financial gain while other hackers seek to earn more information by logging into specific companies to make more malicious attacks against their employees. Consequently, hackers have different and new techniques to fool users such as sending URLs that lookalike like a website for banking or shopping, at the time that the user opens the URL and does transactions, the hacker can steal a lot of critical information like account details, credit card information, users' personal information, passwords, and identity theft.

URL phishing is a cyber-attack that uses URLs and e-mails as a technique to trick the users to believe that the URL or e-mail is a trustworthy person or business in electronic communication, such as a note from their company or a request from their bank, for instance, to download the attachment or to click a link. At that moment, attackers gain the access to the users together with their information and Computer. Furthermore, phishing websites or e-mails are designed to mimic the look of a real company webpage/email [1].

The rapid evolution of intelligent techniques such as artificial intelligence (AI) including Machine Learning (ML) and Deep Learning (DL) are effective in providing security for the operations of computing and cybersecurity management.

The Variety of AI characteristics, from detecting and extrapolating patterns, providing security to adapting to a new environment make it to be a pivotal part of technological systems such as computer vision, and cybersecurity.

To perform feature extraction and selection in classic machine learning techniques, human expertise is needed. Feature selection and classification tasks are separated. In order to optimize the models' performance, deep learning fills that gap using a single phase for detection and classification. deep learning models eliminate the need for manual feature engineering and dependence on third-party services due to automatic learning and feature extraction, unlike machine learning. Moreover, high performance and end-to-end problem solving are the major advantages of deep learning over traditional machine learning techniques, especially with cases of large datasets [2,3].

It is not easy to choose the appropriate method for a certain application. Selecting an inappropriate algorithm or method would result in unpredicted outcomes, resulting in a waste of effort and eventually affecting the model's accuracy and efficiency [4], especially when phishers keep changing their attacking techniques to leverage the systems' vulnerabilities and the users' unawareness. As a result, a variety of anti-phishing technologies have been created to detect phishing threats early on and safeguard users. To deal with evolving phishing assaults, security techniques based on deep learning mechanisms have become increasingly prominent [5,6]. However, selecting an effective phishing detection model with an appropriate dataset that would yield excellent results remains a challenge.

Deep learning applications are used in different industries such as automated driving, facial recognition, and medical devices to name a few. Deep learning instructs computers to mimic human brains through learning by example. Further, in deep learning, a computer model learns to perform classification tasks directly from the complex dataset as text, sound, and images. Deep learning models can accomplish state-of-the-art accuracy, sometimes the accuracy exceeds human performance as well. Deep learning models require large amounts of labeled data for training, substantial computing power, and neural network architectures that contain many layers [7].

The robustness of the deep learning algorithms has encouraged the proposal of many methods for dealing with phishing websites from extracting features to classifying URLs. Numerous methods that assist in detecting phishing attacks were applied by using different, new, and known features such as URL length, frequency of keywords, lexical features, and incorporating features such as works listed in table 1.

As a result, we were motivated to find a solution to the phishing issue effectively using deep learning. This paper adopted an empirical approach to explore the performance of the two techniques LSTM and CNN with one dataset and produced great results. The final goal of this paper was to classify whether the URL was phishing or legitimate by using LSTM, CNN, and integration of both.

We propose a phishing detection system based on deep learning approaches to classify whether the URLs are phishing or legitimate. The proposed system is helpful in information security and cybersecurity research domains, and deep learning-based detection and classification systems. Our work provides insights into the effectiveness of using LSTM and CNN for the purpose of classification of phishing URLs to prevent financial damages and cyber-crimes.

The contribution of the proposed work is listed as follows:

• The LSTM, CNN, and integration of both are proposed in this work to predict the outcome possibilities of URL.

• The prediction results of the proposed LSTM, CNN, and integration of both are evaluated for the accuracy rate, precision, and recall.

• This work uses deep learning-based techniques due to its ability to handle huge data, extract features then classify URLs automatically and efficiently.

• This work considers a dataset with 80 features with the highest accuracy LSTM-CNN algorithm of 94%, then the CNN achieved 92% while LSTM. Achieved 91%.

• We highlighted several limitations based on the findings of the previous studies and recommended possible solutions to address these issues.

The rest of this paper is organized as follows: Section 2 presents a literature review. Section 3 discusses our proposed solution along with its methodology. Experimental results and discussion are the content of Section 4. Limitations and future work were highlighted in Section5.

# Literature Review

The phishing websites problem is complex and is a challenge in itself, because of there is no definitive solution exists to put an end to all the threats effectively. To address phishing websites, much deep learning-based phishing website detection solutions have arisen. Moreover, deep learning has become more promising in cyber-security. In this section, several previous works which use deep learning approaches for phishing detection websites were shown in table1 summarizes.

## LSTM

SU Yang [8] presents a new method that uses the LSTM Recurrent Neural Networks (RCNS) algorithm for detecting phishing attacks that adopts the LSTM deep learning method and optimizes the training method of the model in combination with the characteristics of RNN. The main advantages of using LSTM are its ability to capture large amounts of data and its ability to automatically learn complex features. This solves the tricky problem for other machine learning methods. The datasets used are from yahoo and PhishTank. This work showed an accuracy of 99.1%.

## CNN

Model-based on deep learning proposed in [9], it utilized a character-level CNN to detect the phishing URL. The dataset contained four datasets from PhishTank, from common- crawl and common- crawl and Alexa. The study was conducted by implementing the system of phishing detection by using CNN at a character level to learn the URL’s sequential information, then max or average pooling was applied to determine important features, and the pooled features were then passed to fully connected layers for classification. The stochastic gradient descent algorithm (SGD) is used to train the network. In addition to the system not being required to reach the content of the website or network, it used sequential pattern attributes for the fast classification of the exciting URL. The results show that the suggested model attained an accuracy of 95.02% on the dataset and an accuracy of 98.58%, 95.46%, and 95.22% on benchmark datasets which perform better than the current phishing URL models. Compared to the various machine and deep learning algorithms.

Shweta Singh et al. [10] presented a phishing detection system implemented using deep learning techniques to prevent attacks. The dataset contained 37,175 phishing URLs and 36,400 legitimate ones. The study was conducted by applying CNN. Further to the advantages of this system, no feature engineering is required since the CNN extracts features from the URLs automatically through its hidden layers. The framework is that the input text passed through the embedding layer, and a matrix was created and passed to CNN. The results showed that the proposed system achieved an accuracy of 98.00% compared to the previous models.

Due to a new kind of phishing which is called two-dimensional code phishing attacks, The authors of [11], fend off this threat by proposing a relative detection method. The data was collected from the FlickrLogos-32 dataset which is a publicly available logo dataset that contains 32 different logo brands. The study was conducted by enhancing the traditional approach which is an improved Feature Pyramid Network (FPN) combined with a Faster R-CNN logo recognition method. Logo extraction, logo recognition, and identification are the three main processes of the system. Logo extraction extracts logo images from two-dimensional code. Based on the extracted logo, recognition of the logo through Faster R-CNN and identifying the logo. The final process which is identification is about evaluating the consistency of the logo between the truly identified and its described identity. The results showed the efficacy of the method in logo recognition, which can be used for two-dimensional code phishing attack detection in comparison to the other logo recognition methods and phishing detection methods.

Chidimma Opara et al.[12] proposed HTMLPhish, deep learning that depends on the data-driven end-to-end automatic phishing web page classification method. The data was collected from HTML documents using a web crawler and the dataset contained more than 50,000 HTML documents comprehensive dataset of HTML contents were presented in a real-world distribution. To learn appropriate feature representations from the HTML document, HTMLPhish used CNNs to learn the semantic dependencies in the textual contents of the HTML. Moreover, they implemented convolutions on a concatenation of the matrix of character and word embeddings to guarantee the effective embedding of new words in the test HTML documents. This method could examine context features from HTML documents without considering extensive manual feature engineering. The results showed that HTML- Phish is obtainable over 93% Accuracy which indicates a temporally stable result.

Due to security vulnerabilities and cyber threats to internet users. Artificial intelligence-based algorithms through machine learning and deep learning techniques are developed, such as [13]. The work aimed to implement a phishing detection system to prevent the cyberattacks before they reach users by using a CNN with n-gram features that are extracted from URLs, determining which of the n-gram feature extraction techniques are more effective, and discovering what parameters of the n-gram work best. The best results are achieved with single characters. Through experiments, it was discovered that training the model with 70 characters gives better results with a time of around 34 seconds for 1 epoch for training. As a result, the designed system can classify a URL in 0.008 seconds. The algorithm could reach an 88.90% accuracy level which is quite successful with the High-Risk URL dataset.

Texception is a novel deep learning architecture [14] that takes a URL as an input and predicts whether it belongs to a phishing attack. Whereas The URL of a web page that hosts the attack provides a substantial source of information to determine the maliciousness of the web server. Texception is different from classical approaches since it uses both character-level and word-level information from the incoming URL depending less on manually crafted features or feature engineering. Texception uses multiple parallel convolutional layers and can grow deeper or wider, this flexibility enables Texception to generalize better for new URLs using the Microsoft SmartScreen service dataset. The results of production data showed that Texception outperforms previously proposed deep learning models for detecting phishing website URLs by increasing the true positive rate by 126.7% at an extremely low false-positive rate (0.01%) which is crucial for our model’s healthy operation on the internet scale.

Since individuals and organizations worldwide are becoming increasingly exposed to various cyberattacks due to the faster growth of phishing websites, the improvement of cyber defense and effective phishing detection is required. Suleiman Y. Yerima et al. [15] proposed a deep learning model based on 1D CNN that utilizes CNN for high accuracy classification to distinguish phishing sites from genuine sites. According to the results, the model was evaluated using 6,157 genuine and 4,898 phishing websites as a dataset. The model is perfect to detect new, previously unseen phishing websites. Furthermore, the model gained 98.2% as a phishing detection rate with an F1-score of 0.976.

## Integration of LSTM and CNN

Nguyet Quang Do et al. [16] concentrated on analyzing the performance of different deep learning algorithms to detect phishing to aid organizations in choosing and adopting suitable solutions based on their technological needs. The data was obtained from the University of California Irvine Machine Learning Repository (UCI)dataset, which contained 11055 phishing and benign URLs with 30 website features. They utilized various deep learning algorithms, which comprise DNN, CNN, LSTM, and gated recurrent unit (GRU). Due to finding the optimal parameter to gain great accuracy, it was tested that each neural network has different architectures for each of deep learning algorithms. The results demonstrated that a deep learning algorithm does not exist deep learning algorithm that gains the best measure of overall performance metrics.

Image classification and natural language can both benefit from deep learning approaches. In [17] study, an intelligent Phishing Detection System (IPDS) was deployed to explore the possibility of differentiating phishing URLs from unique legitimate URLs by using CNN and LSTM to build a hybrid classification model. Around one million legitimate and phishing URLs were used as a dataset collected from PhishTank and Common Crawl were used. To build the IPDS, The CNN and LSTM classifier used over 10,000 images and one million URLs for training. The sensitivity of IPDS was determined by considering several factors such as number of misclassifications, the type of feature, and split issues. IPDS gained 93.28% as the accuracy of classification.

Many phishing detection methods are computationally expensive and difficult to update their detection rules based on changes in attack patterns. PhishTrim was proposed in [18], which is a lightweight phishing URLs detection method based on deep representation learning, which is fast and adaptive. Skip-gram pretraining model was used to get the initial embedding representation of the URLs. However, Bi-LSTM is used to extract context dependency to further learn the deep representation of URLs. By using CNN, the local n-gram features are extracted and PhishTrim dataset is used. PhishTrim performs better on large-scale datasets with 99.797% accuracy and the ability to detect zero-day phishing attacks.

As a result of the increase of using e-shopping and e-banking, hackers can steal users' personal information and critical details through different ways by disguising themselves as reputed websites. Yazhmozhi. V.M et al. [19] proposed an anti-phishing system to protect the users. The system uses an ensemble model which uses both LSTM and CNN with 2,00,000 URLs as data set taken from PhishTank, virus total, and using Yandex search API. The proposed system performs well with 97% precision and 96% accuracy. The model was deployed with a straightforward UI and can be used in web browsers as a plug-in component.

After a comprehensive literature review, Phishing detection research is a challenging task since phishers are rapidly developing efficient ways to corrupt the current detectors. Research has phishing detection approaches that can be categorized depending on their input such as URL, email, visual screenshot, logos, and HTML content. In terms of URL as input research, most of the studies have proved that URL features such as URL length, characters, frequency of keywords, and frequency of auspicious symbols play well on the datasets collected from virus total, PhishTank, OpenPish, and other open phishing platforms. The results of these studies showed accuracy reaching 90% and more due to using deep learning methods mainly DNN, CNN, and LSTM. On the other hand, some studies use small datasets which affect the accuracy of the proposed systems. Further, some studies used the same deep learning method for feature extraction and classification with different accuracy. In addition to that, the training time was long. Based on that, there is a need for a system that can help to detect phishing URLs efficiently and effectively. Deep learning has attracted increased interest newly due to its performance, and ability to learn the features instantaneously without any manual feature engineering. Under those promises, we use Deep Learning to detect phishing URLs using LSTM, CNN, and integration of both to show their performances in detecting phishing URLs.

Table 1 Comparative Analysis of Literature Review.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ref. | Brief problem statement | Methodology based on DL | Feature Extraction Method | Classification Method | Dataset | Training Instances | Testing Instances | Performance Measures |
| [8] | Designed a new detection system for phishing websites using LSTM Recurrent Neural Networks (RNN). | LSTM Keras and  RNN |  | LSTM  With sigmoid neuron. | 2000 legitimate websites collected from Yahoo Directory and 2,000 phishing websites collected from Phishtank. | 70% | 30% | Accuracy of 99.1%. |
| [9] | Detect phishing without requiring essential manual feature engineering  and requiring prior knowledge about phishing. | A fast deep learning-based solution model. | The proposed model based on the features extracted from URL does not need manually designed hand-crafted features, and it is independent of network accessing. | (Naïve Bayes, Logistic Regression, random forest, XGBoost, and deep neural networks). | Four datasets from PhishTank, from common-crawl and common-crawl and Alexa. | Randomly split. |  | Accuracy, precision, recall, F1-Score, AUC value, training time, and test time. |
| [10] | Design solution phishing detection system. | CNN | CNN extract features from the URLs automatically through its hidden layers. | CNN | High risk URLs dataset. | Phishing URLs and 36,400 legitimate URLs. |  | F-1 score precision, recall, accuracy. |
| [11] | In response to the new threat called two-dimensional code phishing attacks. | improved Faster R-CNN. | The heuristic-based approach. | SVM classification. | FlickrLogos-32 dataset. | 10 training images. | 30 verification images, and 30 test images. | Precision, Recall and F1-measure. |
| [12] | Proposed techniques to fight phishing attacks. | Propose HTMLPhish, a deep learning-based data- driven end-to-end automatic phishing web page classification approach. | Approach can learn context features from HTML documents without requiring extensive manual feature engineering. | CNNs | HTML documents using a web crawler. | 23,000 legitimate URLs and 2,300 phishing URLs. |  | The precision, True Positive Rate, and F-1 score metrics  Area under the ROC Curve (AUC) and the receiver operating characteristic (ROC). |
| [13] | To implement a phishing detection system by using a Convolutional Neural Network with n-gram features that are extracted from URLs.  moreover, determining which of n-gram is more effective and discovering what parameters of the n-gram work best | Deep Learning  n-gram method  CNN model. | N-gram. | Deep Learning-Based. | High-Risk URL. | 85% | 15% | Accuracy 88.90% and 0.008 seconds as run-time efficiency. |
| [14] | Propose a novel deep learning architecture, Texception, that takes a URL as input and predicts whether it belongs to a phishing attack. | CNN, URLNet model and LR. |  | Binary Cross Entropy loss function along with SGD optimize. | Collected by the Microsoft SmartScreen service.  Microsoft’s anonymized browsing telemetry data. | The first two weeks. |  | Accuracy increasing the true positive rate by 126.7% at an extremely low false positive rate (0.01%). |
| [15] | Present a deep learning-based approach to enable high accuracy detection of phishing sites. | CNN | CNN | CNN | UCI dataset | 90% | 10% | Accuracy of phishing detection rate 98.2%. |
| [16] | Analyze the performance of various deep learning algorithms in detecting phishing activities. | Four different deep learning algorithms, including DNN, CNN, LSTM, and GRU. | Traditional neural network. | CNN | UCI dataset | 80% | 20% | The confusion matrix.  The four common measurements (ACC, PR, RC, F1), other metrics, including FPR, FNR, and AUC. |
| [17] | Design and development of a deep learning-based phishing detection solution that leveraged the universal resource locator and website content such as images, text, and frames. | CNN and LSTM algorithm. | Feature extractor algorithm. | IPDS (CNN+ LSTM)  Knowledge mode. | One million URLs collected from both PhishTank and Common Crawl. | 70% of images. | 30% of images. | Classifier prediction performance and training time.  accuracy of 93.28% and 25.4 seconds for time. |
| [18] | Propose PhishTrim, a light- weight phishing URLs detection method based on deep repre- sentation learning, which is fast and adaptive. | Skip-gram pre-training model , Bi-LSTM and  CNN.  “cnn for feature extraction lstm later | CNN | Multiple convolution structures (fully connected layer for phishing URL classification adopts sigmoid activation function. | phishTrim. |  |  | 99.797% accuracy. |
| [19] | Proposed an anti-phishing system to protect the users against phishing. | LSTM , CNN and  RNN |  | LSTM Cross entropy loss function. | Collected dataset around 1,94,800 URLs. from Phish Tank , virus total and using Yandex search API. | 80% | 20% | Accuracy of 96% and the precision is 97% . |

# Methodology

Detecting phishing in URLs is an important aspect of cybersecurity. Commonly, many phishing URLs appear as legitimate URLs to the users because of the attackers. As a result, Attackers can gain access to the personal information of the users that can be misused. This paper proposed a phishing detection system for detecting phishing URLs. In order to detect URL phishing and show the robustness of the system. Further, the system was implemented by using two different techniques the following sections describe the methodology used, datasets preparation, deep learning approaches, and the training and testing in detail.

## Proposed System

In this section, we present the details of the proposed model configuration. The framework of the proposed model consists of four stages. The first stage is about the features of the URLs which obtained from [20], the second stage is about the data that processing which contained detect null values and scaling values of features then each feature will select that contributes most to the target variable by using SelectKBest, the third stage is training the model using LSTM, CNN and LSTM-CNN by building the deep learning model then compile it and finally the evaluation of the model by using a set of performance metrics were measured to assess the performance of the model in detecting phishing websites, the fourth stage is the classification of the webpage URL either to legitimate or phishing.

Diagram

Description automatically generated

Figure 1 Framework of the System.

## Dataset Preparation and Preprocessing

Data collection plays an essential role in terms of research validity and reliability. In our approach, before implementation, appropriate and consistent data are needed so the system’s results are acceptable. The phishing dataset is taken for the predictive analysis in this research work. After prepossessing the dataset containing the URL features, with 6723 records of 80 features, so many features in the first dataset therefore the SelectKBest method is used with the value of the best 30 features in it. As shown in Fig (2), it has 51 legitimate URLs and 49 phishing URLs.The considered dataset was cleaned using the data preprocessing which includes deciding null values in addition to scaling each feature to a given range using the MinMaxScaler method. Then the resulted dataset has been considered for several numbers of experiments over the LSTM, the CNN, and LSTM-CNN.

Chart, pie chart

Description automatically generated

Figure 2 Phishing Dataset.

## Training and Testing

The dataset was divided into 20% testing and 80% as training. One of the factors affecting the performance of deep learning algorithms is the selection of hyperparameters during training. Hyperparameter values ​​can be optimized to improve the performance accuracy of phishing detection models. These parameters include the number of layers, the number of neurons in each layer, the batch size, the learning rate, the dropout rate, and the number of periods, to name a few [21]. Choosing an appropriate number for each parameter will enhance the LSTM, CNN, and LSTM-CNN model’s performance, so each value is selected based on the value that enhanced the performance. One of the main parameters of the system is the age, which is considered the number of iterations of training after the deep learning model is built and compiled, its value set to 50 epochs.

Table 2 Training and testing dataset distribution.

|  |  |  |
| --- | --- | --- |
| **Dataset distribution** | **phishing** | **Benign** |
| **#URLs for training** | 2420 | 1613 |
| **#URLs for testing** | 1594 | 1096 |

## Deep Learning Approaches

Deep learning, a sub-branch of machine learning, has received a lot of attention in the previous decade. Recent advances in processing power and increased data storage capacities have greatly aided the applicability of deep learning methods. Deep learning-based models have so produced state-of-the-art outcomes on huge datasets for a variety of challenges, including Image processing, natural language processing and machine translation which show the best results. Moreover, the task of phishing URL classification was also used with deep learning systems, with encouraging results [22]. Different classification methods are applied to detect phishing websites and then evaluated by different performance metrics. The classifiers evaluated in this research work are LSTM, CNN, and LSTM-CNN. Convolutional layers are characterized by their ability to extract useful knowledge and learn the internal representation of time-series data, while LSTM networks are effective for identifying short-term and long-term dependencies. Based on the experimental results that show the hybrid model with great results in terms of performance. Therefore, we will limit ourselves to the description of the CNN-LSTM hybrid model.

* **LSTM :** Long Short-term Memory is an adaptive recurrent neural network (RNN), which is a type of recurrent neural network where each neuron is swapped by a memory cell which is additional to the conservative neuron on behalf of an internal state. The layers of LSTM compromise of memory blocks which repeatedly linked blocks, each block contains one or more recurrently connected memory cells.

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 overlooked, as well as an output gate that prohibits or enables the inner state to be viewed from the outside [23]. LSTM has been shown to be an effective strategy for detecting phishing URLs. [24,25]

The workflow of the LSTM for classifying the URL starts after loading, preprocessing, and splitting the dataset. LSTM model starts with the first layer which is the input layer that uses the 79-length vector, and then the LSTM layer which has 128 neurons and will work as the memory unit of the model. After LSTM, the dense layer, which is an output layer with a sigmoid function, sigmoid function helps in providing the labels.

* **CNN: It’s** a convolutional neural network (CNN), which is a discriminative architecture that performs well in processing two-dimensional data with grid topologies, such as images and videos. In terms of time delay, the CNN concept outperforms the NN concept. The weights are shared in a temporal dimension in the CNN idea, which reduces calculation time. The standard NN's generic matrix multiplication is thus replaced in the CNN. As a result, the CNN technique minimizes the weights, lowering the network's complexity [23].

The workflow of the CNN for classifying the URL starts with the first step by fetching the labeled training data of the URLs, then randomly split into train and test sets. After we prepared the training and test data finally the data was trained by making the architecture of the CNN including the input, output, and layers. After each convolution, we add a max-pool layer to extract the most significant elements in each convolution and turn them into a feature vector.

Next, we add dropout regularization to ensure that that model does not overfit. In the end, the model classifies the resulting output of this layer using a sigmoid function.

* **LSTM-CNN:** A model is made up of CNN layers that extract features from input data and LSTM layers that predict sequences [26]. Furthermore, a study [27] found that combining a 1D convolution layer and an LSTM layer improves the accuracy of malicious URL identification when compared to models that exclusively use LSTM layers. As a result, when constructing the system, this paper chose 1D convolutional and LSTM architecture to train the URL features.

The workflow of the CNN-LSTM, after preprocessing the dataset, the dataset is split the into train and test sets, the data normalize before feeding into the model, then the model is implemented which consists of the CNN layer, LSTM layer, in addition to the dense layer to avoid overfitting on the dataset, and finally, the model classifies the resulting of the output layer using a sigmoid function.

Diagram

Description automatically generated

Figure 3 The architecture of LSTM-CNN.

# Evaluation and results

## Evaluation metrics

This section summarizes the metrics used to measure the result of the deep learning approaches. Typically, the performance of the machine learning prediction algorithms is measured by using some metrics based on the classification algorithm. In this paper, the prediction results are evaluated by using the metrics such as Precision, Recall Confusion matrix, and Accuracy of the system were used to estimate the system [28].

**Precision:** The precision of the prediction algorithm is the number of correctly phishing webpages correctly that is belonging to the actual phishing webpages.

1

**Recall:** Recall of the prediction algorithm is the number of correctly predicted phishing URL made of all URLs in the dataset. It is a true positive rate.

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 pro- duces four different results, such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

**F1-Score:** Is the process of taking the harmonic mean of a classifier's precision and recall. It can be combined into a single metric.

3

4

**Confusion matrix:** It produces prediction results in the matrix form with the information of the number of correctly phishing URLs, incorrectly predicted phishing URLs, and errors of incorrect, and correct prediction phishing URLs.

## Results

For the experimental results, we calculate the Accuracy, Precision, Recall, F1 score, and confusion matrix of prediction algorithms. Since in most of the prediction models research papers', the accuracy of the prediction model has been considered as one of the common performance metrics, the evaluation of the proposed system was performed based on the prediction accuracy of the approaches presented in this paper in section 3. We use a dataset that consists of 15368 records of 80 features. Then we preprocessed it by detecting null values, scaling values of features, and then selecting features using SelectKBest, we trained the LSTM, CNN, and LSTM-CNN classifiers by building and compiling the model and finally we evaluated the system.

The three proposed methods showed a great result shown in table1, which reflects the good choice of the value of the suitable parameters. After implementing, training, and testing the LSTM, CNN, and LSTM-CNN techniques, the results showed some level of improvement in phishing detection through LSTM-CNN algorithm since it has the highest accuracy as 94%, followed by CNN algorithm which achieved 92%, while LSTM achieved 91% of prediction accuracy as figure4 illustrated.

Table 3 The performance results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evaluation Metric** | **LSTM** | **CNN** | **LSTM-CNN** |
| **Accuracy** | 0.914870 | 0.952416 | 0.947955 |
| **Precision** | 0.849282 | 0.918595 | 0.918474 |
| **Recall** | 0.963801 | 0.970136 | 0.958371 |
| F1-score | 0.902925 | 0.943662 | 0.937998 |

Chart, bar chart

Description automatically generated

Figure 4 Evaluation metrics.

For the LSTM, figure 5 illustrates the confusion matrix of the LSTM algorithm. The x-axis represents the percentage of predicted values, and the y-axis represents the percentage of true values. It can be seen that the LSTM algorithm predicts 1065 (true positive) of the deceased cases correctly, with 40(false positive) of misclassification. The LSTM algorithm produces the highest value of 0. 91.

Chart, treemap chart

Description automatically generated

Figure 5 Confusion matrix of LSTM.

For the CNN, as shown in figure 6 illustrates the confusion matrix of the LSTM algorithm. The x-axis represents the percentage of predicted values, and the y-axis represents the percentage of true values. It can be seen that the LSTM algorithm predicts 1059 (true positive) of the deceased cases correctly, with 46(false positive) of misclassification. The LSTM algorithm produces the highest value of 0. 92.

Chart

Description automatically generated

Figure 6 Confusion Matrix of CNN.

Figure 7 illustrates the confusion matrix of the LSTM-CNN algorithm. The x-axis represents the percentage of predicted values, and the y-axis represents the percentage of true values. It can be seen that the LSTM-CNN algorithm predicts 1072 (true positive) of the deceased cases correctly with 33 (false positive) of misclassification. The LSTM-CNN algorithm produces the highest value of 0. 89.

Chart

Description automatically generated

Figure 7 Confusion matrix of LSTM-CNN.

After analyzing the result, we can conclude that the CNN-LSTM algorithm outperforms LSTM and the CNN algorithms in the detection of phishing.

# Comparison with existing approaches

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ref.** | **Proposed Methodology**  Table 4 Comparison of existing approaches. | **Dataset** | **Advantages/ disadvantages** | **Accuracy** |
| 16 | They utilized various deep learning algorithms, which comprise DNN, CNN, LSTM, and gated recurrent unit (GRU). | 11055 phishing and benign URLs with 30 website features. | **Advantages**   1. It will build a secure connection between a Mail user agent and a mail transfer agent. 2. The classification process is fast. 3. Many new features are created.   **Disadvantages**   1. It takes more time for parameters selection as well in network learning. 2. When the size of the training set increases, accuracy decreases. 3. It has no standard classifier and is higher in cost. 4. It requires high memory. | 96.70% |
| 17 | An intelligent Phishing Detection System (IPDS) was deployed to explore the possibility of differentiating phishing URLs from unique legitimate URLs by using CNN and LSTM to build a hybrid classification model. | One million legitimate and phishing URLs were used as a dataset. | **Advantages**   1. The combination of both CNN and LSTM was used to resolve the problem of a large data set and higher classifier prediction performance. 2. Combining the LSTM and CNN algorithms gives better results in the detection of a phishing website.   **Disadvantages**   1. Many drawbacks of this scheme are still unsolved. 2. Its accuracy is less compared to other methods or techniques. | 93.28% |
| 18 | PhishTrim was proposed, which is a lightweight phishing URLs detection method based on deep representation learning, which is fast and adaptive. | PhishTrim dataset is used. | **Advantage**   1. New phishing websites can be easily detected. 2. It only needs a URL for phishing detection.   **Disadvantages**   1. Not all phishing websites are detected 2. The blacklist website list is uploaded regularly which takes time. | 99.797% |
| 19 | Proposed anti-phishing system to protect the users. The system used both LSTM and CNN. | 2,00,000 URLs as data set taken from PhishTank, virus total, and using Yandex search API. | **Advantages**   1. It gives good performances with very high precision. 2. It is less time-consuming. 3. The work in the method can be extended by using bidirectional.   **Disadvantages**   1. It has insufficient parameter selection techniques. 2. It does manual parameter tuning. | 96% |
| 29 | They trained the CNN and CNN-LSTM models to detect phishing URLs. | They created a dataset from four different sources: MalwareDomainlist , MalwareDomain for malware URLs, PhishTank and OpenPhish for phishing URLs. | **Advantages**   1. This method offers a diverse combination of CNN and RNN. 2. Machine learning and deep learning-based malicious URL detection can foreclose detection systems built using blacklisting and regular expression method.   **Disadvantages**   1. The detection of malicious websites is a task that needs continuous development. So, it is time-consuming. 2. Deep neural network architecture is very complex therefore, understanding the background mechanics of a neural network model remains a black box. | 98% |
| 30 | CNN-LSTM to effectively identify malicious URLs and improve the accuracy rate. | The dataset is from PhishTank. | **Advantages**   1. Its algorithm is attention build. Attention is a mechanism that can flexibly select context information. 2. This method has better recognition and detection effect than traditional method.   **Disadvantages**   1. Due to the high rate of the false judgment of the blacklist method and the difficulty of updating the heuristic rules. | 98.18% |
| 31 | They used CNN to detect phishing webpages by automatically extracting URL features. The method is based on CNN and  Bi-LSTM. | URLs  dataset. | **Advantages**   1. LSTM only considers the forward information and does not consider the backward information. This issue, however, can be resolved in bidirectional LSTM. 2. It has caught much attention among researchers, and some of their research works in the phishing detection domain.   **Disadvantages**   1. Time Series forecasting problems. 2. It is hard to predict in categories close to the number of products. 3. There is no standard and holistic guideline for selecting these hyperparameters to achieve the highest performance accuracy. | 96 % |

# Limitations and Future Work

After testing and evaluating our proposed system, we can see that the system outperforms existing works and showed excellent results. However, the proposed system has some drawbacks. First is regarding the training time, it takes a long time to train since the epoch is 50, but on the other hand, it is better than other models in terms of accuracy. Another drawback is that the model does not check the status of the URL of the website i.e., is the website active or not which impacts the results. To overcome these limitations and drawbacks, our intention in future work is to enhance the training process in terms of time and improve the feature engineering in order to enhance the training process in terms of accuracy. Besides, check the website’s status. Finally, we plan to experiment with relatively new deep learning algorithms and compare different deep learning algorithms' performance for detecting phishing websites.

# Conclusion

Recently, the improvement of the Internet and technologies have had a significant impact on increasing online purchases and transactions. On the other hand, online transactions lead to unauthorized access to the sensitive information of users as individuals or enterprises. Security is the most important aspect to protect internet users from stealing their information while they are communicating through internet applications. Besides, security is a challenging task in the internet domain. Phishing is one of the known attacks that gain the users' information through an URL that looks identical to the actual webpage URL. Detecting phishing attacks plays a significant role to prevent attackers from gaining access to the users' information. As there is a growth in the number of victims due to inefficient security technologies exists, an intelligent technique is needed to protect users from cyber-attacks. With the rapid development of deep learning techniques, deep learning has proven valuable development compared to traditional signature-based and classic machine-learning-based solutions due to its high performance and end-to-end problem-solving. In this work, the LSTM, CNN, and LSTM-CNN algorithms are used to detect and classify the URLs of the websites as either phishing or legitimate. Based on the evaluation of the proposed system, the detection of phishing websites accomplished excellent results. The used deep learning algorithms applied on the same dataset vary in their performance. The LSTM-CNN algorithm outperforms CNN and LSTM in terms of accuracy which reaches 94%, while CNN and LSTM achieved 92%, and 91% as accuracy respectively.

**Refernces:**

[1] "Phishing | Phishing Techniques", Phishing.org, 2022. [Online]. Available: https://www.phishing.org/phishing-techniques. [Accessed: 21- Apr- 2022].

[2] "Why Deep Learning over Traditional Machine Learning?", Medium, 2022. [Online]. Available: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063. [Accessed: 22- Jun- 2022]

[3] R. Grover, "Deep Learning - Overview, Practical Examples, Popular Algorithms | Analytics Steps", Analyticssteps.com, 2022. [Online]. Available: https://www.analyticssteps.com/blogs/deep-learning-overview-practical-examples-popular-algorithms. [Accessed: 22- Jun- 2022]

[4] Sarker, I.H. Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective. SN Comput. Sci. 2021, 2, 154.

[5] Feng, J.; Zou, L.; Nan, T. A Phishing Webpage Detection Method Based on Stacked Autoencoder and Correlation Coefficients. J.Comput. Inf. Technol. 2019, 27.

[6] Y. Huang, Q. Yang, J. Qin and W. Wen, "Phishing URL Detection via CNN and Attention-Based Hierarchical RNN," 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2019, pp. 112-119, doi: 10.1109/TrustCom/BigDataSE.2019.00024.

[7] "What is Deep Learning and How Does It Work?", SearchEnterpriseAI, 2022. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network. [Accessed: 23- Jun- 2022]

[8] Y. Su, "Research on Website Phishing Detection Based on LSTM RNN," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020, pp. 284-288, doi: 10.1109/ITNEC48623.2020.9084799.

[9] A. Aljofey, Q. Jiang, Q. Qu, M. Huang and J. Niyigena, "An Effective Phishing Detection Model Based on Character Level Convolutional Neural Network from URL", Electronics, vol. 9, no. 9, p. 1514, 2020.

[10] S. Singh, M. P. Singh and R. Pandey, "Phishing Detection from URLs Using Deep Learning Approach," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277459.

[11] W. Yao, Y. Ding and X. Li, "Deep Learning for Phishing Detection," 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom), 2018, pp. 645-650, doi: 10.1109/BDCloud.2018.00099.

[12] C. Opara, B. Wei and Y. Chen, "HTMLPhish: Enabling Phishing Web Page Detection by Applying Deep Learning Techniques on HTML Analysis," 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207707.

[13] M. Korkmaz, E. Kocyigit, O. K. Sahingoz and B. Diri, "Phishing Web Page Detection Using N-gram Features Extracted From URLs," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1-6, doi: 10.1109/HORA52670.2021.9461378.

[14] F. Tajaddodianfar, J. W. Stokes and A. Gururajan, "Texception: A Character/Word-Level Deep Learning Model for Phishing URL Detection," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2857-2861, doi: 10.1109/ICASSP40776.2020.9053670.

[15] S. Y. Yerima and M. K. Alzaylaee, "High Accuracy Phishing Detection Based on Convolutional Neural Networks," 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), 2020, pp. 1-6, doi: 10.1109/ICCAIS48893.2020.9096869.

[16] N. Do, A. Selamat, O. Krejcar, T. Yokoi and H. Fujita, "Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical Study", Applied Sciences, vol. 11, no. 19, p. 9210, 2021.

[17] M. Adebowale, K. Lwin and M. Hossain, "Intelligent phishing detection scheme using deep learning algorithms", Journal of Enterprise Information Management, 2020.

[18] L. Zhang and P. Zhang, "PhishTrim: Fast and adaptive phishing detection based on deep representation learning," 2020 IEEE International Conference on Web Services (ICWS), 2020, pp. 176-180, doi: 10.1109/ICWS49710.2020.00030.

[19] Y. V. M, B. Janet and S. Reddy, "Anti-phishing System using LSTM and CNN," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-5, doi: 10.1109/INOCON50539.2020.9298298.

[20] "URL 2016 | Datasets | Research | Canadian Institute for Cybersecurity | UNB", Unb.ca, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/url-2016.html. [Accessed: 28- Nov- 2020]

[21] S. Mahdavifar and A. Ghorbani, "Application of deep learning to cybersecurity: A survey", Neurocomputing, vol. 347, pp. 149-176, 2019.

[22] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, “Deep learning in computer vision: A critical review of emerging techniques and application scenarios,” Machine Learning with Applications, vol. 6, p. 100134, Dec. 2021, doi: 10.1016/j.mlwa.2021.100134.

[23] M. A. Adebowale, K. T. Lwin and M. A. Hossain, "Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection," 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2019, pp. 1-8, doi: 10.1109/SKIMA47702.2019.8982427.

[24] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas and F. A. González, "Classifying phishing URLs using recurrent neural networks," 2017 APWG Symposium on Electronic Crime Research (eCrime), 2017, pp. 1-8, doi: 10.1109/ECRIME.2017.7945048.

[25] W. Chen, W. Zhang and Y. Su, "Phishing detection research based on LSTM recurrent neural network", Proc. Int. Conf. Pioneering Comput. Sci., pp. 638-645, 2018.

[26] S. Ariyadasa, S. Fernando and S. Fernando, "Detecting phishing attacks using a combined model of LSTM and CNN", International Journal of ADVANCED AND APPLIED SCIENCES, vol. 7, no. 7, pp. 56-67, 2020.

[27] T. Pham, V. Hoang and T. Ha, "Exploring Efficiency of Character-level Convolution Neuron Network and Long Short Term Memory on Malicious URL Detection", Proceedings of the 2018 VII International Conference on Network, Communication and Computing - ICNCC 2018, 2018.

[28] V. Lakshmi and M. Vijaya, "Efficient prediction of phishing websites using supervised learning algorithms", Procedia Engineering, vol. 30, pp. 798-805, 2012.

[29] A. Vazhayil, R. Vinayakumar, and K. P. Soman, “Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks,” in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Jul. 2018, pp. 1–6. doi: 10.1109/ICCCNT.2018.8494159.

[30] “MALICIOUS URL RECOGNITION AND DETECTION USING ATTENTION-BASED CNN-LSTM -KSII Transactions on Internet and Information Systems (TIIS) | Korea Science.” https://www.koreascience.or.kr/article/JAKO201905959996575.page (accessed Jun. 20, 2022).

[31] Q. Zhang, Y. Bu, B. Chen, S. Zhang, and X. Lu, “Research on phishing webpage detection technology based on CNN-BiLSTM algorithm,” J. Phys.: Conf. Ser., vol. 1738, no. 1, p. 012131, Jan. 2021, doi: 10.1088/1742-6596/1738/1/012131.