International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

Design of Automated Website Phishing Detection using Sequential Mechanism of RCL Algorithm

C. Rajeswary[1], **M. Thirumaran**[2]
[1]Department of Computer Science and Engineering,
Puducherry Technological University
Pillaichavady,Puducherry-605014, India
rajeswary.c@pec.edu
[2]Department of Computer Science and Engineering,
Puducherry Technological University
Pillaichavady,Puducherry-605014, India
thirumaranptuniv@edu.in

Abstract— The phishing outbreaks in internet has become a major problem in web safety in recent years. The phishers will be stealing crucial economic data regarding the web user to perform economic break-in. In order to predict phishing websites, many blacklist-based phishing website recognition methods are used in this study. Traditional methods of detecting phishing websites rely on static features and rule-based schemes, which can be evaded by attackers. Recently, Deep Learning (DL) and Machine Learning (ML) models are employed for automated website phishing detection. With this motivation, this study develops an automated website phishing detection using the sequential mechanism of RCL algorithm. The proposed model employs Long-Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Random Forest (RF) models for the detection of attacks in the URLs and webpages by the similarity measurement of the decoy contents. The proposed model involves three major components namely, RF for URL phishing detection, CNN based phishing webpage detection, and LSTM based website classification (i.e., legitimate and phishing). The experimental result analysis of the RCL technique is tested on the benchmark dataset of Alexa and PhishTank. A comprehensive comparison study highlighted that the RCL algorithm accomplishes enhanced phishing detection performance over other existing techniques in terms of distinct evaluation metrics.

Keywords- Cybersecurity; Phishing detection; Websites; Machine learning; Deep learning; Hybrid models.

I. INTRODUCTION

The term phishing came from the term fishing [11]. It started in 1990s with the outbreak called the American Online (AOL) and is still a major challenge, taking the first position in the landscape of cyber threat [8]. [31] This is a kind of cyberattack that happens in the internet and includes implementing ingenious and fraudulent processes for obtaining web user's personal data and economic login details.

By employing verified addresses to send fraudulent emails, the frauds gather the credentials of the imprudent victims by fooling them. After opening an email, it will lead the customers to top fake pages that trick the attackers into giving financial data, for example, user login and passwords. Specialized deception sends malware onto the PCs to steal the user credentials. Economic advancement is the prime target for these outbreaks and is human intelligence is greatly associated to the phishing. These days, most of the phishing outbreaks happen in e-commerce, news, games, social media, weather report, financial institution, logistics, and cryptocurrency. This pose a major threat as new unique websites are getting increased in the present days. Hence, it becomes difficult and challenging for the software professionals in detecting phishing outbreaks effectively.

In 2018, a group called Kimsuky in Korea involved in phishing and their target was not just the finance but also the Korean security associated foreign institutions, defence, and diplomatic data [1]. In the pandemic year, COVID-19 attackers have raised their attention to target the vulnerable victim by applying various tricks based on human fear, anxiety, and stress factors. They used a lot of tricks to make them explore by sending similar emails from WHO and Health care centers to make the people more vulnerable [1] [2]. Spear fishing is more targeted at open-source intelligence. Spear phishing attackers conduct an intelligent survey on social media before making an attack; if they make an attack, it is tough to find [3].

The statistical data reports that in the Internet Crime Complaint (IC3) in the year 2018, the United States' FBI accused that 2.7 billion dollars were stolen because of the phishing emails through online banking and trading activities. Based on the Anti-Phishing Working Group (APWG) in the year 2019, by employing secure HTTP and Socket Layer (SSL) encrypting, a large number of weak websites are discovered as 266,387 in the First Quarter of July and 182,465 in the Second Quarter and the numbers doubled in the Fourth Quarter [4]. [6] Phishers utilize emotional tricks in reaching the victims which is based on their curiosity towards some buzz words, lucky draw, and gambling. They started to collect and learn from the user, who is exploited by risk-taking behavior decision-making style and make themselves more vulnerable. Those factors are psychological vulnerabilities.

[5] Detecting Phishing activities can be accomplished in two manners: Whitelist Phishing Detection (WPD) and Blacklist Phishing Detection (BPD). WPD contains genuine websites but it is difficult to centralize a global database. Also, this assumes automatic WPD. BPD contains phishing websites and also it is difficult in maintaining a worldwide database to identify emerging sites. The presented model blocks the

***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

distrustful websites and URLs, thus improving the phishing detection and reducing the cybercrime ratio.

[7] This introduces a method for detecting phishing emails. This helps in detecting the emails earlier before it could reach the targeted victim. This model adopts the technique called Feature Engineering (NLP) and ML. [10] The Counterfeiting, Affiliation and Stealing, and Evaluation (CASE) comprises overall and interpretable quaternary factors with anti-phishing statistical factors. This identifies the aspects of the social engineering by CASE reflects the web content factors. Fig. 1 depicts the ML working process based phishing detection.

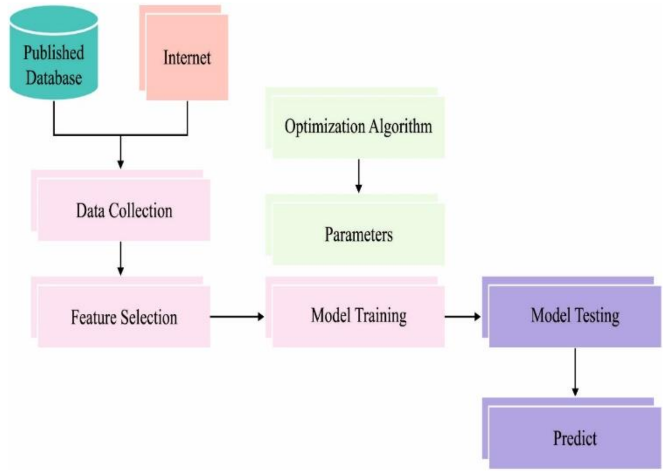


Fig. 1. Working process of ML based phishing detection process

II. PROPOSED WORK

The purpose of the phishing detection and proposed system using an RCL is explained in this section. The RCL algorithm combines the working principle of the RF algorithm, CNN, and LSTM in sequential order. 2.1 Defines the Problem statement. 3.Algorithms and Techniques used for Phishing Detection. 4. Explains the Proposed RCL algorithm for Phishing detection model. 5. Section shows the Experimental Results and Evaluation of Performance Metrics. Section 6 Concludes with the future work.

A. Problem statement

The problem states that the RCL algorithm must detect and classify attacked and non-attack web sites. Consider an input contains the URL and Webpage contents.

$$N = \{w_1, w_2, w_3, \ldots\ldots w_i\} \quad (1)$$

where w_i is webpage i=1,2,3…. n.

$$(x) = \begin{cases} L_{i>1} \\ L_{i<1} \end{cases} \quad (2)$$

Where L_i is more than 1, then it is a Legitimate webpage.
and if L_i is lesser than 1, then it is Phishing webpage.

According to the RCL algorithm, after detection of each phishing webpage, the model consists of three parts, $D_p = \sum_{P=0}^{n} \binom{n}{p} U_p + W_p + M_p$ (3)

Where D_p is the detected page, U_p is the URL, W_p is the webpage contents, and Mp is the Memory page from LSTM.

B. Proposed Automated Phishing Detection Architecture

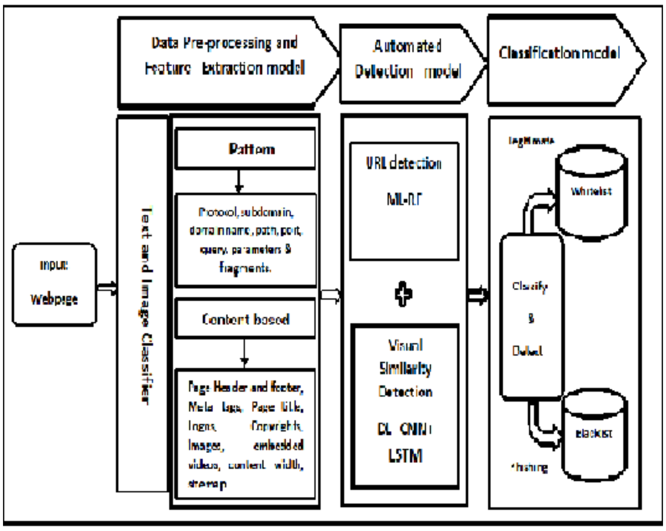

Data Pre-processing and Feature Extraction model
Automated Detection model
Classification model
Input: Webpage
Text and Image Classifier
Pattern
Protocol, subdomain, domain name, path, port, query, parameters & fragments.
Content based
Page Header and footer, Meta tags, Page title, logos, Copyrights, Images, embedded videos, content width, sitemap
URL detection ML-RF
+
Visual Similarity Detection DL CNN+ LSTM
legitimate
Whitelist
Classify & Detect
Blacklist
Phishing

Fig. 2. Proposed framework Phishing Detection.

The proposed framework has three Modules. They are 1. Data Pre-processing and feature extraction Module.2 Automated Detection Module. 3. Classification Module.

This study aims in expanding the detection of phishing ability related to the features of URL and webpage contents such as text, image, frame, logo, and visuals by applying the algorithms like LSTM, RF, and CNN. The input is given from the dataset such as Alexa for Legitimate page and PhishTank for Phishing page. By using the classifiers such as text and image, the URL text such as Protocol, domain name, sub-domain, path, port number, characters, query, parameters, and fragments. The webpage contents such as page header and footer, meta tags, page title, logos, logos, copyrights, images, sitemap, and favicons were extracted and fed into the Hybrid detection model.

It has the RF algorithm to detect URL phishing and CNN and LSTM algorithms to identify the web page content. After detection, it will be fed to the classification model. It performs classification based on the phishing aspects. If 1, it is stored in the blacklist repository; if 0, it is stored in the whitelist repository. [38] White and blacklist are used in list-based phishing detection methods. URLs that have been identified as phishing sites are added to blacklists. By employing blacklists, attackers are prevented from attacking a URL or IP address that is the same through malicious code; the protective framework updates the blacklist. To identify phishing and genuine websites, WPD systems give data about safe and trustworthy web resources. A website that isn't on the whitelist is regarded to be malicious.

***Article Received:** 26 August 2023 **Revised:** 20 October 2023 **Accepted:** 02 November 2023*

C. Data Pre-Processing And Feature Extraction Model Workflow

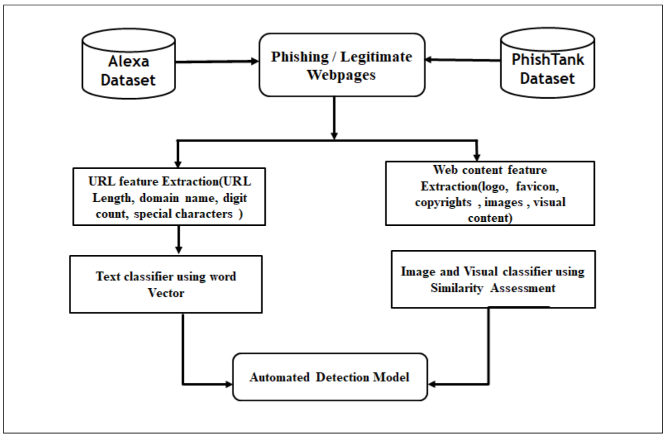
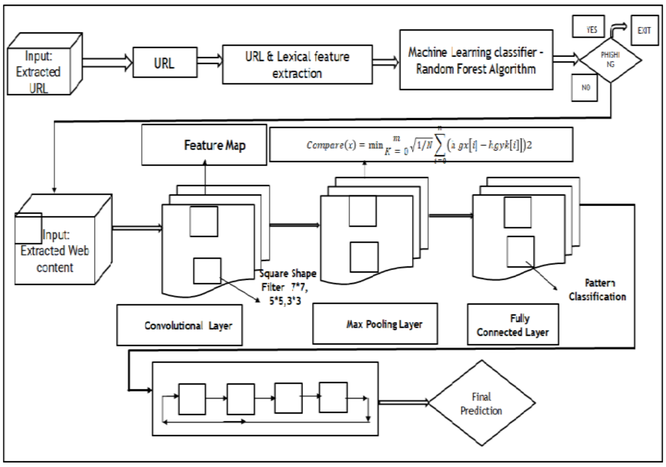

Alexa Dataset
Phishing / Legitimate Webpages
PhishTank Dataset
URL feature Extraction(URL Length, domain name, digit count, special characters)
Web content feature Extraction(logo, favicon, copyrights , images , visual content)
Text classifier using word Vector
Image and Visual classifier using Similarity Assessment
Automated Detection Model

Fig. 3. Data Pre-Processing and Feature Extraction Model

Data is gathered from phishing data from the Alexa and PhishTank datasets. After data gathering from the dataset, the website structures like the size of URL, digit count, text, domain name, and special characters like (@,#,%,^,&,*) are extracted by implementing word vector. Later, web contents like favicon, copyrights, images, logos, and visual content are categorized based on the resemblance evaluation by employing the visual and image classifier. [24] The pivotal role of feature extraction is it aims at extracting the features of popular websites and has more effect on detecting the attacked sites.

D. Working flow of Hybrid Phishing Detection model


Input: Extracted URL
URL
URL & Lexical feature extraction
Machine Learning classifier - Random Forest Algorithm
PHISHI NG
YES
EXIT
NO
Feature Map
$Compare(x) = \min_{K=0}^{m} \sqrt{1/N} \sum_{i=0}^{n} (\grave{a}\, gx[i] - \grave{h}gyk[i])2$
Input: Extracted Web content
Square Shape Filter 7*7, 5*5,3*3
Pattern Classification
Convolutional Layer
Max Pooling Layer
Fully Connected Layer
Final Prediction

. Fig. 4. Working Principle of Automated Detection Model.

The hybridization of ML and DL algorithms can work well at detecting and classifying whether the subpage is phishing or a legitimate one.[8] An advanced approach for detecting URL phishing using lexical features based on RF used the dataset for our experiment as ISCXURL-2006. Different ML algorithms were examined, but RF had the highest accuracy in URL detection (95.57). Hence, we choose the best classifier algorithm detection. It is combined with the CNN and LSTM and thus named the RCL algorithm.

The CNN &LSTM does not require manual feed data for future engineering classification or detection. CNN has three layers: Convol-layer, Maxi-pool layer, and thoroughly on layer. It has to detect the distinctive features in the image. Each layer consists of a feature map. Each layer should see the different shapes and use a set of filters such as RGB filter, square shape filter, and rectangular filter.

- It uses the image comparison formula:
- $Compare(x) = \min_{K=0}^{m} \sqrt{\frac{1}{N}} \sum_{i=0}^{n} (h(g)x[i]) - h(g)yk[i]))^2 \qquad (4)$
- x is denoted as the phishing content to be identified.
- The brand-named favorite icon is represented by y(k) and h(g) specifies the histogram function.
- M denotes the number of images.

It does an automated feed by using the feed-forward and reverse propagation network. It can detect newly generated phishing websites fast and accurately. The current work is very mime efficient because it makes seeing the newly developed webs efficient and thus increases the system's robustness.

E. Classification model

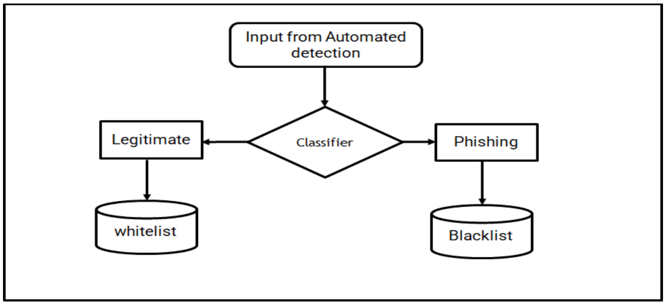


Fig. 5. Classification Model

The input from the automated phishing model is given to the classifier. The information is classified based on the legitimacy and vulnerable values on the web page. If the website is trustworthy, it will be stored in the whitelist repository; if not, it will be labeled as phishing and stored in the blacklist repository.

III. METHODS AND MODELS IMPLEMENTING IN PHISHING DETECTION

A. RF Model:

[26] RF model in ML performs better on decision-making ability because of its strong branching decision trees.[11] This is implemented for the classification of the chosen subset from training dataset using classification and regression. After classification, the decisions are aggregated, and the average of them not only for controlling the overfitting but also improves the accuracy.

As decision trees are implemented for creating RFs, it is significant to comprehend them before understanding RF. To create a root set, adding the factors has the major influence on the label to the factor. These may be calculated by employing various scoring models like the Gini Index. After root selection, the rest characteristics were investigated and compared. The characteristics were graded, and the most significant ones were used to determine the root [40].

B. LSTM Algorithm

[30] Hochreiter and Scmidhuber introduced the LSTM architecture and its application in various fields. It can learn long-term dependencies; it has built-in mechanisms. Its feed-forward network ability decides how to control information and must be memorized or abandoned for feature use. [23] Like the RNN, the LSTM network also has a chain-like topology and a cell structure called the gate. The gate such as input-output-forget. These gates make predictions over the information

flow.[27] However, the RNN has only one hidden layer and one state to remember the short-term input.[30]. The complex and dynamic information is abstracted and filtered using internal memory.[28] It is ideally suited for identifying the detailed patterns in time-series data; it significantly reduces the gap length in remembering the exiting patterns.[29]

RF	**CNN**	**LSTM**
Let U1 and U2 be the dataset	X-input image	i_t - input gate O_t - output gate
x-URL factors like (text, HTTP, domain name, subdomain, special characters, etc.)	M-No. of images h(g) and y(k) are histograms	C_t - saved input h_t - current output σ_g Sigmoid σ_c = tanh w - weight matrix b - bias
M-feature set	RELU-Rectifier Linear Unit	f_t - forget gate
E(t)-Entropy of two test data	w-weight b-bias	X_t - the current input
E (t, x) = Entropy of feature x	Conv1, Conv2 - convolutional layer1, 2 respectively.	C_t - saved input C_{t-1} - status of the last saved page

Table 1. Nomenclature:

C.CNN MODEL:

[25] The CNN model is under the unsupervised learning model in DL. In this, the prior data is studied and then apply the studied data for creating new data. It performs an automatic recognition of recently created phishing websites. [29] This model has biases parameters and weights for learning data like other neural networking's. It is majorly implemented for image recognition and categorization. Images are sent into each convolutional layer's neuron in the format Height * Width * Channel [19], where the dimensions of the pictures are measures, and the channel count is the channel number. Additionally, it has various kernels, designated as k, whose Height -1 +1 image is created by convolving the kernel with the picture. Then the average pooling determines the feature map average, whereas the max pooling layer chooses the most crucial data from a collection of feature maps(n). With sequential input, the feature, the map is sent to the fully connected layer.

IV. PROPOSED RCL ALGORITHM FOR PHISHING DETECTION

A. RF ALGORITHM FOR URL PHISHING DETECTION

```
Input: Webpage- Alexa U1 and PhishTank U2 dataset.
Output: Detect & Classify Phishing and Legitimate.
Start
1. Elect URL features U1 randomly.
2. For every x in M, where x is {x1.x2, x3……xn}
    i) Do the Information gain Calculation:    IG (t, x) = E(t)-E (t, x)
            (5)
    ii) For computing the entropy among two testing data.
            E(t) = $\sum_{i=1}^{c} - pi \log_2 pi$
            (6)
    iii) For computing the entropy of features of x
            E(t,x) = $\sum c\epsilon x \; P(c)E(c)$
            (7)
    iv) Choose node d which has high data gain
    v) Create the sub-node from the rrot nodes.
    vi)To build a tree, repeat steps i, ii, iii, till the minimum sample numbers are need to divide.
3. Repeat Steps 1 & 2 for N number of times for building the format of N trees.
4. If (x>0)
exit ()
else
Go to step 5:
```

RF Pseudocode [31]

B. CNN ALGORITHM FOR PHISHING IMAGE DETECTION

```
Input: Image as X
Output: Detected Phishing images
Start
5. Initialization of CNN function with X images
        a.    Represent bias and weight.

6. The initial image is pre-processed via implementing the shape function, [Pixel, Pixel_ Y SLICE COUNT])
7. It is then fed to the conv1 for transformation by utilizing RELU_ACTIVATION_FUNC (CONV3D (X, w [0]) +b [0])
8. DROPOUT () the image size.
9. The dropout image is fed to the MAX_POOL3D from the convo layer.
10. It relates the image after dropping out and reshaping using the following:    $compare(x) = \min_{K=0}^{m} \sqrt{1/N} \sum_{i=0}^{n} (h\,gx[i] - hgyk[i]])2$
11. It raises the value of bias and weight to 1 in Convolution 2 layer. It accomplishes the rectifier function: conv2←RELU_ACTIVATION_FUNC (Conv3D (Conv1, w[1] + b[1]))
12. DROPOUT () the size of the image
13. The dropout image is later forwarded to the fully connected (FC) layer.
14. It accomplishes the inverse function by FC← RESHAPE (Conv2, INVERSE (weights [2]))
15. It increases the value of bias and weight as 2 by employing FC← RELU_ACTIVATION_FUNC (FC × weight [2] + biased [2])
16. DROPOUT () the size of the image
17. The resulted value of the output is raised to 3.
18. OUTPUT← Fully_Connect * w [3] + b [3]
19. Display the detected image
20. Jump to Step 21.
End function.
```

Pseudocode for CNN [33]

C. LSTM MODEL FOR PHISHING

```
Input: Input detection samples from CNN
Output: Phishing pages in Memory gate.
Begin
21. Initialization of input samples to $A_0$.
22. Perform detection: if (i>0)
Legitimate
else
Phishing
23. Determine each neuron's output value in advance by calculating the parameters such as:
 $f_t$, $i_t$, $c_t$, $o_t$, $h_t$, $c'_t$.
    a.  $f_t = \sigma_g (w_f \times x_t \times u_f \times h_{t1} + b_f)$
            (8)
    b.  $i_t = \sigma_g (w_i \times x_t \times u_i \times h_{t-1} + b_i)$
            (9)
    c.  $o_t = \sigma_g (w_o \times x_t \times u_o \times h_{t1} + b_o)$
            (10)
    d.  $c'_t = \sigma_c (w_c \times x_t \times u_c \times h_{t-1} + b_c)$
            (11)
    e.  $c_t = f_t \times c_{t-1} + i_t \times c'_t$
            (12)
    f.  $h_t = o_t \times \sigma_c (c_t)$
            (13)
 24. Compute the error value using loss function.
 25: Once detecting the Webpage, the phishing/authentic parameters are updated.
```

Input/Output equation for LSTM [34].

V. EXPERIMENTAL RESULTS AND EVALUATION OF PERFORMANCE METRICS:

A. EXPERIMENTATION ALGORITHMS

The automated Phishing detection approach is analyzed with the assistance of three diverse models such as the ML and DL. i.e., RF model for detecting the URL, and CNN and LSTM for web page content.

[40] [31] By utilizing the dataset to construct a bootstrapped dataset, a RF was initially created. While samples from the old dataset would still be present in the new dataset, they would have been randomly chosen and added to the new table, making it possible for some samples to appear more than once in the new table.

The decision trees' root was created in the second stage, which involved choosing a randomly selected subset of the characteristics to examine using the scoring method. Another random subset from the remaining characteristics was once again chosen to add children to the root, and the following child was then chosen after analysis.

Incorporating the ML and DL for enhancing the efficiency of the phishing detection and later more precisely recognizing phishing sites, i.e., during model training, characteristics from a hidden layer outcome of the CNN approach are utilized as the input value for training the LSTM. The suggested approach was validated [33][34][41].

B. PROPOSED MODEL'S IMPLEMENTATION

The presented automated RCL model's implementation is employed by utilizing the i5 core processor with the configuration of 8 GB RAM, 3.4 GHz, and a 2GB graphics card. The language utilized for the implemented and execution is Python and Anaconda 3. These days, Python is the majorly used for ML processes as it contains a huge libraries like NumPY, Scikitlearn, etc, and also Matplotlib is of great assistance in data dispersion and for plotting the analyzed values in a pictorial format. The hybrid algorithm RF-CNN-LSTM efficiently detects the URL and webpage phishing accurately. Two datasets were used as input, such as Alexa and PhishTank Dataset. Figure 7. and Figure 8. Shows the performance evaluated for Accuracy, Recall, Precision, and F1-Score in URL detection. Figure 9 and Figure 10. Directs the performance evaluated for Accuracy, Recall, Precision, and F1-Score n Webpage phishing.

The datasets and metrics for evaluating the detection performance are introduced in this section. The experiment is carried out conducted with the existing dataset [10][20][19][18], along with the recently used phishing datasets. First, we take the URL dataset to train with the RF classifier algorithm more accurately. The evaluation metrics will measure the accuracy of the detection, recall, precision, F1- score, precision, and FPR, respectively. After evaluating the URL, the web page features are evaluated using CNN and LSTM.

C. DATASETS

We used two datasets such as Alexa and PhishTank. For the Alexa dataset www.Kaggle.com and PhishTank www.phishtank.org, those two websites provide the free phishing datasets on URL and Website accordingly. In Alexa, out of 507195, 72% are good, and 28% are bad [35] and in PhishTank, out of 7179486, 3019506 are phished sites are valid [36.] The datasets were used for utilizing the achievement of the presented detection system using RF, CNN, and LSTM approaches and for enhancing the detection rate of the website' phishing activities.

D. PERFORMANCE METRICS

The achievement of the models is measured by analyzing the metrics like precision, accuracy rate, recall, true positive, true negative, F1-score, error rate, detection time, false positive, and false-positive rate.

- **True Positive (TP-**α**)** - for precisely anticipated phishing sites.
- **True Negative** (**TN-**β**)** - for precisely anticipated legitimate sites.

  False Positive (FP-(∂)**) and False Negative (FN**-(μ)) - once the improper categorization occurs.
- **Accuracy (A)** - It represents the phishing percentage and genuine URLs and the successfully detected websites by the classifier.
- **Accuracy (%)** - $A = \frac{\alpha+\beta}{\alpha + \beta+\partial+\mu} \times 100$

  (14)
- **Recall (R)** - The recall value is expected to be 1, which indicates that the overall detection has been completed and if it is 0, then the detection is expected to be less.

  Recall - $R = \frac{\alpha}{\alpha + \mu}$

  (15)
- **Precision (P) -** The value of precision is the higher

  Table 2. Performance Analyzed for URL Phishing

Article Received: *26 August 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *02 November 2023*

- detection's percentage. The classifier's better achievement is computed based on the value of the higher precision.
- **Precision** - $P = \frac{\alpha}{\alpha + \partial}$ (16)
- **F1 score (F1)** - This is computed based on the average Precision and Recall detection rate.

  F1 - Score $F1 = \frac{2 \times P \times R}{P \times R}$ (17)
- **FPR** - Good detection shows a low FPR value.

  FPR - $\boldsymbol{FP}R = \frac{\partial}{\partial + \beta}$ (18)
- **Error rate:** It is calculated in terms of the percentage of the average value obtained between the True positive and negative with a truly positive and false positive.

  Error rate (%) - $E = 100 - \frac{\alpha + \beta}{\alpha + \partial}$ (19)
- **Detection rate (%):** The percentage of phishing detection without error.

  Detection rate (%): $D = \sum_{i=0}^{n} \frac{\mathbf{D1}(x,y) - \boldsymbol{D2}(x,y)}{\text{Total no of detection}} \times 100$ (20)

Table 3. Performance Analyzed for the Phishing of the Websites

The approaches utilized for testing and training	ACCURACY	PRECISION	F1-SCORE	RECALL
RF	0.944	0.944	0.954	0.952
CNN1	0.985	0.982	0.985	0.989
LSTM	0.987	0.985	0.987	0.992
CNN2	0.997	0.995	0.997	0.999
CNN-LSTM	0.997	0.999	0.998	1.000
RCL (Proposed Algorithm)	**0.998**	**1.000**	**0.999**	**1.000**

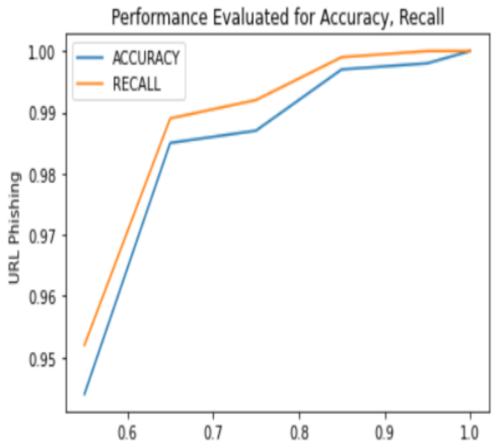


Fig 6. Performance Metrics analyzed for Accuracy and Recall - URL detection

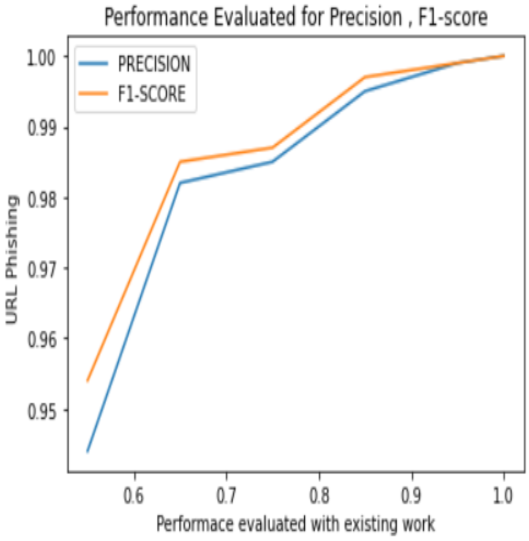


Fig.7. Performance Metrics analyzed for Precision, and F1-Score - URL detection

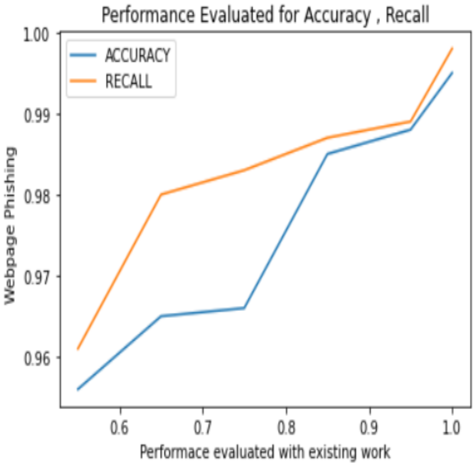


Fig.8. Performance metrics evaluated for Accuracy, Recall -Webpage Detection

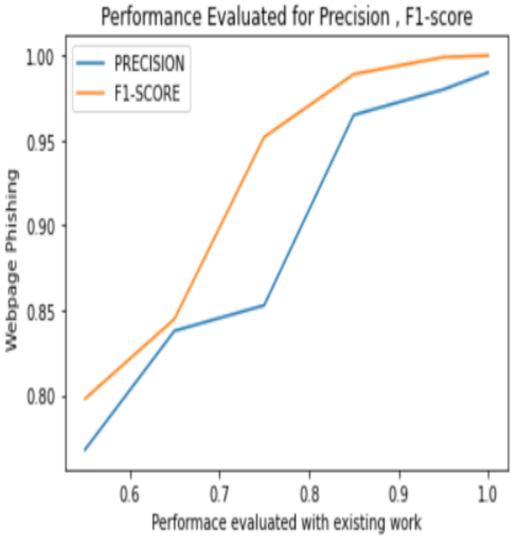


Fig.9. Performance metrics evaluated for Precision and F1-Score - Webpage Detection

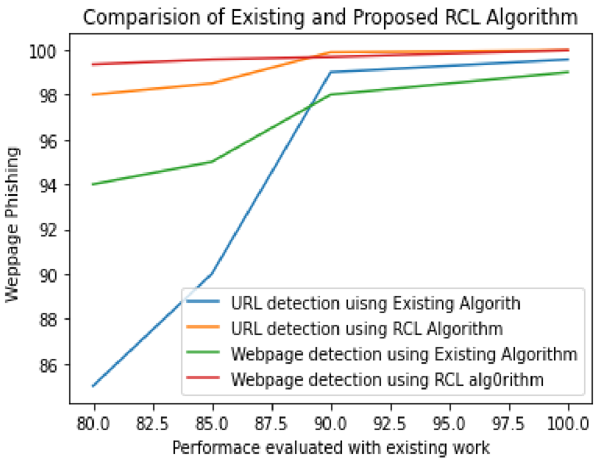


Fig.10. Overall Comparison of Existing and Proposed RCL Algorithm

VI.CONCLUSION AND FUTURE WORK

To detect the complete features of web page contents with satisfying metrics like precision, recall, accuracy, and F1-score. We proposed an automated Phishing Detection based on the RF-CNN-LSTM algorithm. We first introduce the RF algorithm for detecting the URL features from the dataset such as Alexa and PhishTank. After detecting the URL, if it is seen as phishing, it exits, otherwise, it looks for the webpage contents for detection. Sometimes the legitimate URL also has phishing in its webpage content. The unaware content is detected deeply by applying DL models like CNN and LSTM.

The investigational outcomes depicted that the performance metrices like precision, recall, accuracy, FPR, and F1-score are associated with the current datasets. The proposed RCL algorithm works in a sequential process to detect the URL and Web page contents effectively. The presented study accomplishes better precision, recall, accuracy, FPR, and F1-score values. Figure 10. demonstrates the comprehensive achievement of the automated models with the present approaches and its accurate detection in all features. In limitation terms, the proposed work must concentrate more on other attacks on Phishing web pages. Only two datasets were used in our work; phishing can be tested with more datasets and metrics in the future.

REFERENCES

[1] Jaeil Lee, Yong Joon Lee, Dongh wan Lee, Hyukjin Kwon, And Dongkyoo Shin. Classification of Attack Types and Analysis of Attack Methods for Profiling Phishing Mail Attack Groups. IEEE ACCESS, 9, pp.no: 80866-80872, doi:10:1109/Access.2021.308497.

[2] Jan Devos, Geert Poels, And Eric Laermans, Hossein Abroshan. COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempt During the Pandemic. IEEE ACCESS, 9, pp.no: 121916-12129, doi:10:1109/Access.2021.3109091.

[3] Xia Yang Chen, Xing Tong Liu, Lei Zhang, And Chao Jing Tang. Optimal Defense Strategy for Spear-Phishing Attack Based on a Multistage Signaling Game. IEEE ACCESS, 7, 2019, pp.no: 19902-19921, doi:10.1109/Access.2019.2897724.

[4] P.A. Barraclough, G. Fehringer, J. Woodward. Intelligent Cyber-Phishing Detection for Online. ELSEVIER, Computers and Security, pp.no: 1-17, https://doi.org/10/1016/j.cose.2021.10213.

[5] Nureni Ayofe Azeez, Sanjay Misrab, Ihotu Agbo Margaret, Luis Fernandez-Sanz, Shafi'I Muhammad Abdulhamide. Adopting Automated Whitelist Approach for Detecting Phishing Attacks. ELSEVIER, Computers and Security, pp.no: 1-18, https://doi.org/10/1016/j.cose.2021.102328

[6] Hossein Abroshan, Jan Devos, Greet Poels, and Eric Laermans. Phishing Happens Beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of Phishing Process. IEEE ACCESS, 9, pp.no:44928-44949, 2021, doi.10.1109/ACCESS.2021.3066383.

[7] Eder Souza Gualberto, Rafael Timoteo De Sousa, Jr. Hiago Pereira De Brito Vieira, Joao Paulo Carvalho Lustosa Da Costa and Claudio Gottschalg Duque, The Answer Is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering. 8,2020, pp.no:223529 - 223547, doi:10.1109/ACCESS.2020.3043396.

[8] Brij, Gupta, Krishna Yadav, Imran Razzak, Konstantinos Psannis, Arcangelo Castiglione, Xiao Jun Chan. A novel approach for phishing URLs detection using lexical-based machine learning in a real-time environment. ELSEVIER, 175, pp.no:45-57, https://doi.org/10/1016/j.comcom.2021.04.023.

[9] Erzhou Zhu, Yuyang chen, Chengcheng Ye, Xuejun Li, and Feng Liofs-NN. An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. IEEE ACCESS, 7, 2019, pp.no:73271-73284, doi:10:1109/ACCESS.2019.292065.

[10] Dong-JeLiua, Guang-Gang Geng, xiao-Bo Jind, Wei, Wan. An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the actual web environment. ELSEVIER, 110, Computers and Security, https://doi.org/10/1016/j.cose.2021.102421.

[11] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, Banu Diri. Machine learning-based phishing detection from URLs. ELSEVIER, 117, pp.no:345-357, Expert Systems and Applications, https://doi.org/10/1016/j.eswa.2018.09.029.

[12] Waleed Ali and Sharaf Malebary. Particle Swarm Optimization -Based Feature Weighting for Improving Intelligent Phishing Website detection. IEEE ACCESS, 8, pp.no:116766-116780, doi:10.1109/ACCESS.2020.3003569.

[13] Abdelhakim Hannousse, Salima Yahiouchee. Towards benchmark datasets for machine learning-based website phishing detection: An experimental study. ELSEVIER, 104, pp.no:1-17, Engineering Applications of Artificial Intelligence, https://doi.org/10/1016/j.engappai.2021.104347.

[14] Yazan Ahmad Alsariera, Victor Elijah deyemo, Abdullateef Oluwagbemiga Balogun, and Ammar Kareem Alazzawi. AI Meta-learners and Extra -Trees Algorithm for the Detection of Phishing Website. IEEE ACCESS, 8,2020, pp.no: 142532-142542, doi:10.1109/ACCESS.2020.3013699.

[15] Eder S. GualBerto, Rafael T.De Sousa, Jr Thiago P.De B .Vieira, Joao Paulo C.L Da Costa, and Claudio G. Duque, " From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection. IEEE ACCESS, 8, pp.no:76368-76385.

[16] Ayman El Aassal, Shahryar Baki, Avisha Das, and Rakesh M. Verma. An In-Depth Benchmarking and Evaluation of Phishing Detection Research for Security Needs. IEEE ACCESS, 8,2020, pp.no:22170-221192, doi:10.1109/ACCESS.2020.2969780.

[17] Maria Sameen, Kyunghyun Han, and Seong Oun Hwang. PhishHaven-An Efficient Real-Time AI Phishing URLs Detection System. IEEE ACCESS, 8,2020, pp.no: 83425-83443, doi:10.1109/ACCESS.2020.2991403.

[18] Jian Feng, Lianyang Zou, Ou,Ye, and Jingzhou Han. Web2Vec Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning. IEEE ACCESS, 8,2020, pp.no:221214- 221224, doi:10.1109/ACCESS.2020.3043188.

[19] Xiao, Wentao Xiao, Dianyan Zhang, Bin Zhang, Guangwu Huc, Qing Li, Shutao Xia. Phishing websites detection via CNN and multi-headed self-attention on imbalanced datasets. ELSEVIER, Computers & Security, 108, pp.no: 1-14, 102372, https://doi.org/10/1016/j.cose.2021.102372

[20] Wei Wei, Qiao Kec, Jakub Nowak, Marcin Kory Tkowski, Rafal Scherer, Marcin Woziake. Accurate and fast URL phishing detector: A convolutional neural network approach. ELSEVIER, Computers & Security,178, https://doi.org/10/1016/j.comnet.2020.107275

[21] https://www.ic3.gov/Media/Y2019/PSA190910

[22] https://enterprise.verizon.com/resources/reports/dbir/

[23] Ali Al Bataineh, and Devinder Kaur. Immunocomputing- Based approach for Optimizing the Topologies of LSTM Networks. IEEE ACCESS, 9, pp.no:78993-799004, doi:10.1109/ACCESS.2021.3084131.

[24] Waleed Ali, Adel A. Ahmed. IET Information Security Research Article Hybrid Phishing prediction using deep neural networks with a genetic algorithm-based feature selection and weighting. 13, pp.no: 659-669, ISSN 1751-8709, doi:10.0149/iet. ifs.2019.0006.www.ietdl.org.

[25] Moruf Akin Adebowale Khin T.Lwin and M.A Hossian. Intelligent Phishing detection scheme using deep learning algorithms. 1741-0398, pp.no:1-15, doi:10.1108.JEIM-01-2020-0036.

[26] Vamsee Muppavarapu, Archana Rajendran, and Shriram Vasudevan. Phishing Detection using RDF and Random Forests, The International Arab Journal of Information Technology, 15, pp.no: 817-824.

[27] Yong Fang, Cheng Huang, Liang Liu and Min Xue. Research on Malicious JavaScript Detection Technology Based on LSTM. IEEE ACCESS, 6, pp.no:59118-59125, doi:10.1109/ACCESS.2018.2874098.

[28] Milail Mohammed Salim, Sushil Kumar, Singh, Jong Hyuk Park. Securing Smart Cities using LSTM algorithm and lightweight containers against botnet attacks. ELSEVIER 2021, Applied Soft Computing,113, pp.no:1-13, https://doi.org/10/1016/j.asoc.2021.107859

[29] Saaransh Baranwal, Siddhant Khandelwal, Anuja Arora. Deep Learning Convolutional Neural Network for Apple Leaves Disease Detection. International Conference on Sustainable Computing in Science, Technology and Management, 2019, pp.no:260-267.

[30] Wei Li, Denis Mike Becker. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling. ELSEVIER 2021, Energy, 237 pp.no: 1-16, https://doi.org/10/1016/j.energy.2021.121543.

[31] http://www.apwg.org/reports/apwg_trends_report_q3_2021.pdf

[32] https://www.researchgate.net/figure/Randon-Forest-classifier-algorithm

[33] Jaishree Ranganathan, Nikhil Hedge, Allen S. Irudayaraj, Angellina A. Tzacheva, Automatic Detection of Emotions in Twitter Data – A scalable Decision Tree Classification Method. ACM ISBN 123-4567-24-567/08/06, https://doi.org/10.475/123_4

[34] Mohit Dua , Drishti Makhija, P.Y.L.Manasa & Prashant Mishra. A CNN-RNN-LSTM Based Amalgamation for Alzheimer’s Disease. Journal of Medical and Biological Engineering,2020, 40, pp.no:688-706, https://link.springer.com/article.10.1007/s40846-020-00556-1

[35] https://towards.net/p/machine-learning/tutorial-on-lstm-a - computational-perspective-f3417442c2cd

[36] www.kaggle.com

[37] www.phishtank.org

[38] Kang Leng Chiew, Kelvin Sheng Chek, Yong, Choon Lin Tan, A Survey of Phishing attacks: Their Types, Vectors and Technical. Approaches,2018, pp.no:1-59, doi: 10.1016/j.eswa.2018.03.050.