IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

# An Improved Model for Detecting Uniform Resource Locator (URL) using Deep Learning

**Palimote Justice[1], Nkue Dumka[2]**

Department of Computer Science, Rivers State University, Port Harcourt, Nigeria[1,2]

**Abstract:** In this fast growing modern technology driven world, the internet is one of the most important technology not only for individual users but also for organization and online business. Nowadays, there are phishers who steals sensitive information like username, password, credit card, personal data etc. Several researchers have design rule-base system for phishing detection which are credited to help people who cannot understand which Uniform Resource Locator(URL) is real or fake address. This paper concentrates on an improve Model for Detecting Phishing URL using Deep Learning. Object oriented Design Methodology was used for system architecture and structure. The support Vector Machine has been used to extract the actual and the visual link from the domain name system(DNS) and compare the actual link and the visual link if they are same. The system uses the Deep Learning model (Generated Adversarial Network) in tensor flow and Keras framework to classify Website URL dataset containing 3207 URLwebsites, 1037 are Real URL websites and 2137 are fake. The dataset was read from directory using the pandas.read_csv function. The dataset was cleaned to make sure there are no null values present. The results of the test showed accuracy of 99.8% of all input website URL classified as either Fake or Real to verify if it's actually a Fake website or Real Website.

**Keywords**- Machine Learning, Deep Learning, Phishing, Generated Adversarial Network and Uniform Resource Locator.

## 1. INTRODUCTION

Phishing is a social engineering attack in which the attacker attempts to compromise a user's credentials or a system by presenting itself as a real business communication. Most organizations have an online business such as sales of product and services (Liu & Ye, 2001)[1]. Phishing is a powerful technique use to mislead people either by giving a feeling that the site is legitimate or by showing some greedy approaches. The main strategy of phishing sites is to collect your personal information illegally like user ID, password, credit card details, personal data etc. to steal information from users of the site. The total number of phishing attacks for first quarter survey were 263,538. It was also significantly more than the 190,942 seen in 3rd quarter of 2017.The number of unique phishing reports submitted to APWG during 1st 2018 was 262,704, compared to 233,613 in 4th quarter of 2017 and 296,208 in 3rd quarter of 2017. (APWG 201) [2].
This study is focused on identifying phishing URL using deep learning. This paper focused on building an Anti-phishing system on URL features.

## 2. LITERATURE REVIEW

Nkue *et al*. [3] developing an efficient model for detecting phishing URL using machine learning technique. Link Guard algorithm was used to extract real and visual links from the Domain Name System (DNS) and compared the actual link and the visual link to check if the links are the same. Support Vector Machine has been used to train and classify URL into genuine and phishing URL. The system was implemented in Python programming language. Experiments were conducted using two publicly accessible website databases to test the efficiency of the identification of phishing websites. 3200 websites samples were used, 2127 genuine websites and 1036 websites were phishing. The LinkGuard and SVM yielded an accuracy of 98.8%.

Aydin *et al*. [4] proposed a system in Feature extraction and classification of phishing websites based on URL, approaches a system of new methods to extract scalable and basic functions. Data were retrieved from Phish Tank and Google's legal Standardized URLs C# programming and R programming were used to obtain the text properties. The dataset and third party service providers received 133 functions. Correlation-based selection of features (CSF) based on subset and accuracy subset based on feature selection approaches used in the Waikato Environment for Information Analysis (WEKA) tool for feature selection and analysis. For performance appraisal, Naïve Bayes and Sequential Minimal Optimization (SMO) algorithms have been compared and SMO is favored for phishing detection by the author rather than Naïve Bayes (NB).

Karabatak *et al*. [5] proposed system in Performance comparison of classifiers on reduced phishing website dataset, In order to get higher order execution, it recommends feature selection algorithms to minimize the dataset components.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

They were also compared to other classification algorithms for data mining. The phishing website dataset was taken from the machine learning library at the University of California, Irvine from the findings, it is shown that execution is increased by certain classification strategies; some of them decrease execution with a reduced portion. Stochastic Gradient Descent (SGD) Bayesian Network, Lazy. K. Star, Randomizable Filtered Classifier, in order to minimize the phishing dataset, Logistic Model Tree (LMT) and ID3 (Iterative Dichotomiser) are useful and Multilayer Vision, JRip, Component, J48, Random Forest and Random Tree algorithms are not valuable for the reduced phishing dataset. With 27 reduced functions, Lazy. K. Star received 97.58 percent accuracy. This analysis was performed with the aid of WEKA tools.

Rishikesh & Irfan [6]. proposed a system in Phishing Website Detection using Machine Learning Algorithms, by extracting and analyzing various features of legitimate and phishing Uniform Resource Locator, it deals with machine learning technology to detect phishing Uniform Resource Locator. They try to overcome the drawbacks of blacklist and heuristics based methods which cannot detect zero-hour phishing attack, by focusing on machine learning techniques. They receive their legal Uniform Resource Locator dataset from www.alexa.com and the respective illegitimate www.phishtank.com website. The data collection contains a total of 36,711 URLs, including 17058 legal Uniform Resource Locator and 19653 Uniform Resource Locator phishing."The legitimate Uniform Resource Locator is labeled "0" and the Uniform Resource Locator for phishing is labeled "1. For detecting phishing websites, Decision Tree, Random Forest and Help Vector Machine algorithms are used. Using the random forest algorithm, which had the lowest false positive percentage, they obtained 97.14 percent precision. Their outcome also illustrates that as they use more data as training data, the classifiers have improved output. However, together with the blacklisting method, they were unable to obtain a hybrid alternative using the random forest.

Yue & Zhang [7]. proposed CANTINA Approach which distinguishes phishing in sites, Based on the TF-IDF data calculation. Saloon can be utilized to look at the substance of a website page to decide if it is authentic or a phishing webpage. The term recurrence (TF) is basically the occasion when a given term shows up in a particular archive. The reverse report recurrence (IDF) is a proportion of the real significance of a term. Obviously, the IDF gauges how regular a term is over a whole assortment of archives.

## 3. MATERIALS AND METHOD

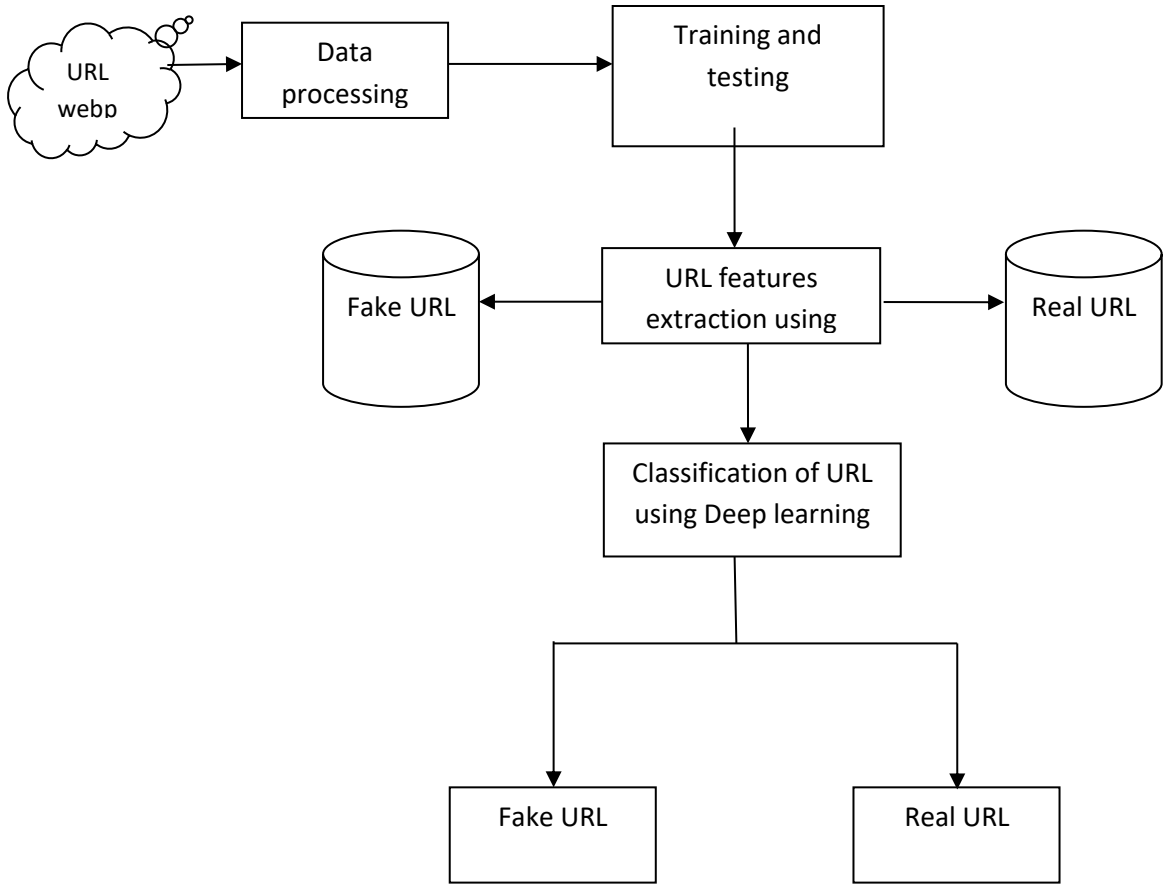


Figure: 1 Architecture of the proposed system design

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

The proposed system includes two Machine learning algorithm. The SVM which is used to extract the URL features from the database of the fake and real URL and compare if they the same, deep learning classify the URL as real or fake.

**Deep Neural Network**

Deep neural network containshidden layer between input and output layer to make classification. Deep neural network in this study is Generated Adversarial Network (GAN). GAN is a neural network that contains generator and discriminator to make classification.

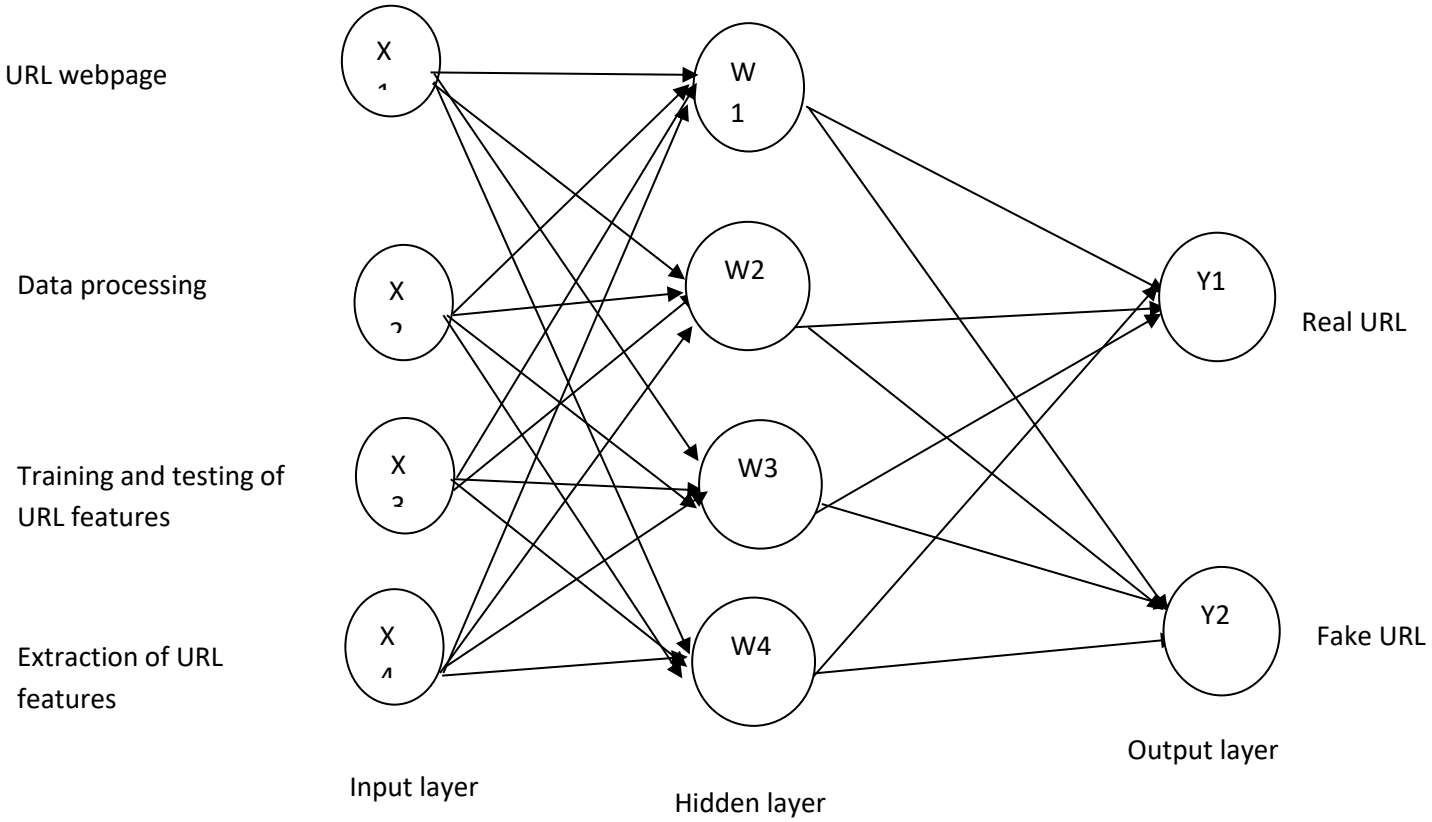


Figure: 2 Deep Neural network

**Algorithm for Support Vector Machine**

I. Import the dataset
II. Explore the data to figure out what they look like
III. Preprocess the data
IV. Divide the data into training and testing sets
V. Extract the URL features
VI. Split the data into attribute and label
VII. Compute attribute values
VIII. If attribute a_ is inputted
IX. V_link1 =decode(v_link);
X. a_link 1=decode(a_link);
XI. return SVM(v_link1,a_link1);
XII. if v_is null
XIII. return analyze v_(a_link),
XIV. end

**Algorithm of Deep Learning**

I. For each training iteration
II. Take a random sample from the training set, label it as x

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

III. Take random sample from fake URL training set, label it as y
IV. Compute the attribute valves
V. End for
VI. If attribute x present value=1
VII. Then deep learning return real URL
VIII. End if
IX. Else if attribute y absent value =-1
X. The deep learning return fake URL
XI. End if
XII. Select attribute x and y
XIII. Compute threshold value for attribute x and y
XIV. Find range value
XV. Select attribute to get threshold value
XVI. Classify fake and real URL by updating the deep learning fake and real URL. Minimize classification error.

## 4. EXPERIMENT

The system uses the Deep Learning model (Generated Adversarial Network) in tensor flow and keras framework to classify Website URL dataset containing 3,207 URL websites, 1,037 are Real URL websites and 2,137 are fake. The dataset was cleaned making sure that there are no null values, duplicate value present. CountVectorizer was used in transforming the URL text to a vector of token counts for a better use as input to the Deep Learning Algorithm. This dataset was further divided into x and y variables where x contains the input (which are the URL Website) and y contains the output (which will display either a Real or Fake URL Website) by using the train_test_split module from sklearn.model_selction library. The Deep learning model was built with a total of three dense layer with takes in 8672 inputs and 1 output, a batch size (batch size equals the total dataset thus making the iteration and epochs values equivalents) of 20 and epoch (The training steps) value of 30. Figure 3 shows the dataset of the original URL website. Figure 4 shows that all the unwanted and null values have been remove, therefore, making the dataset to be balance for training the deep learning model in order to have a better training performance and yield a batter result. Figure 5 shown correlation metrics is use to estimate the linear relationship between the returns of multiple asset. The Dataset showing in figure 6 contains 1,037 are Real URL websites and 2,137 are fake.

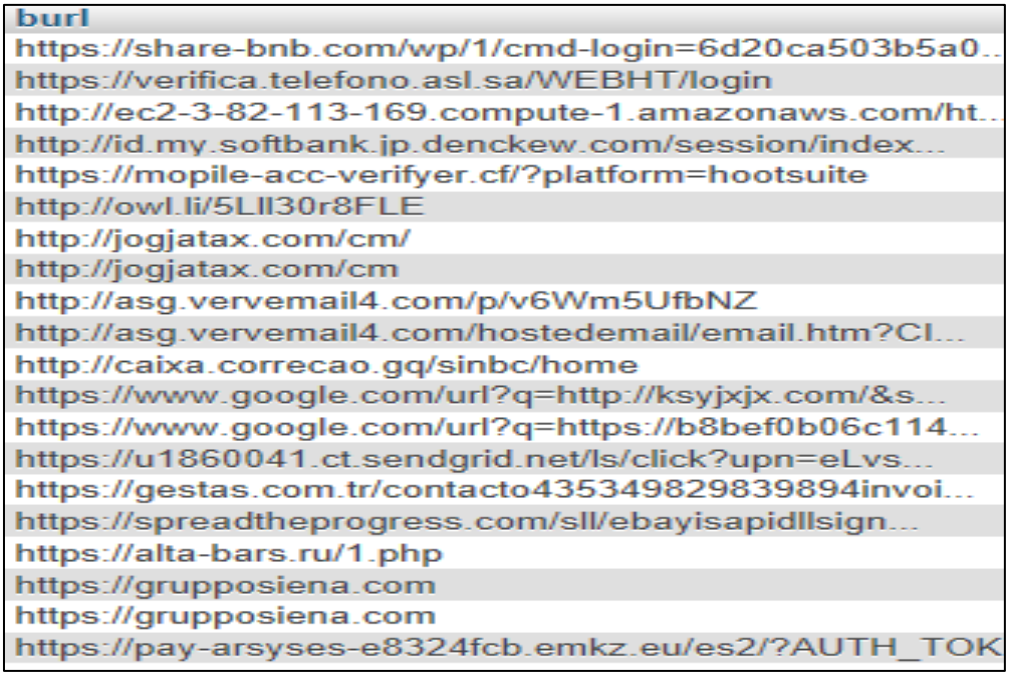


Figure 3: Display of fake the Dataset URLS

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

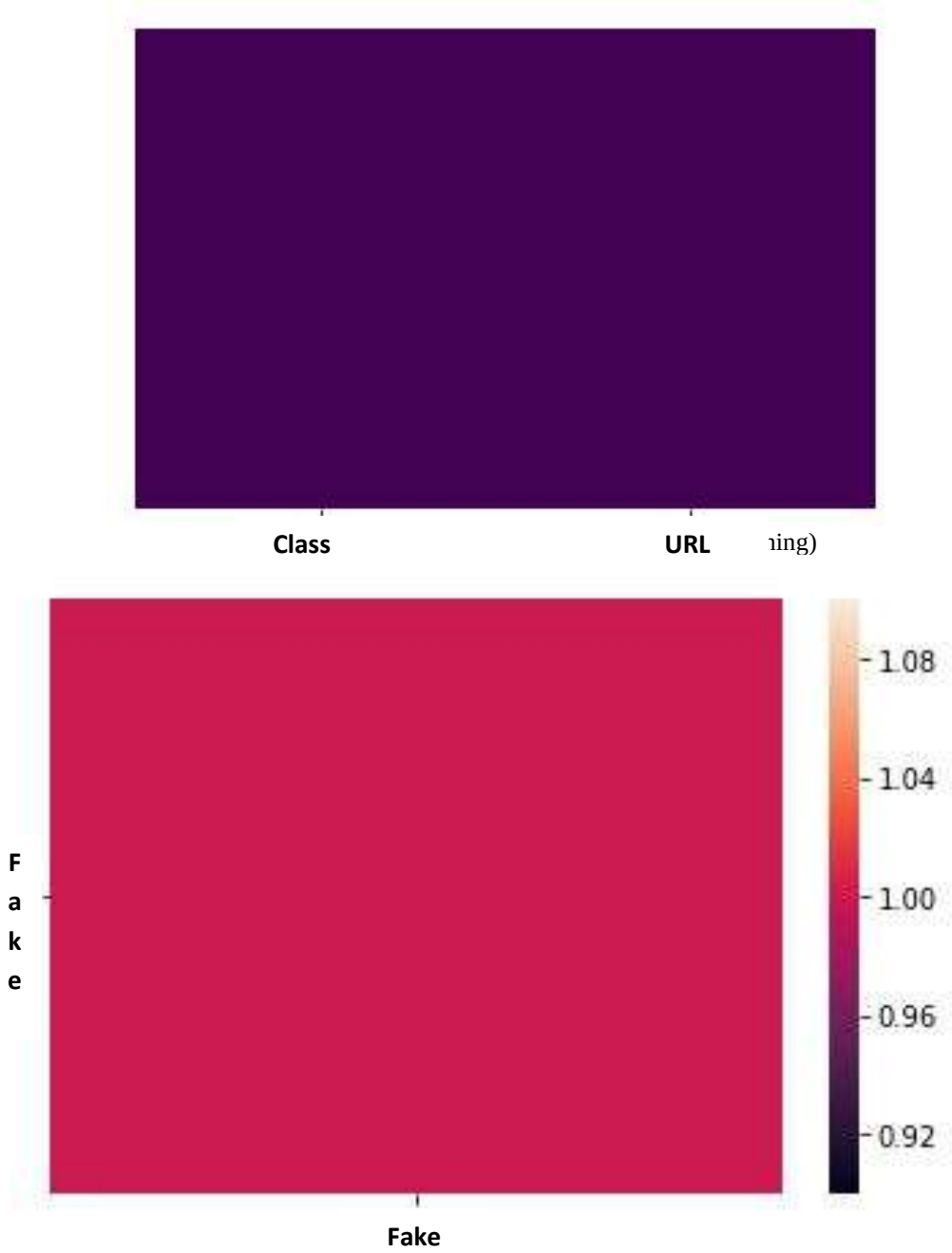


Figure 5: Correlation metrics of the dataset

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

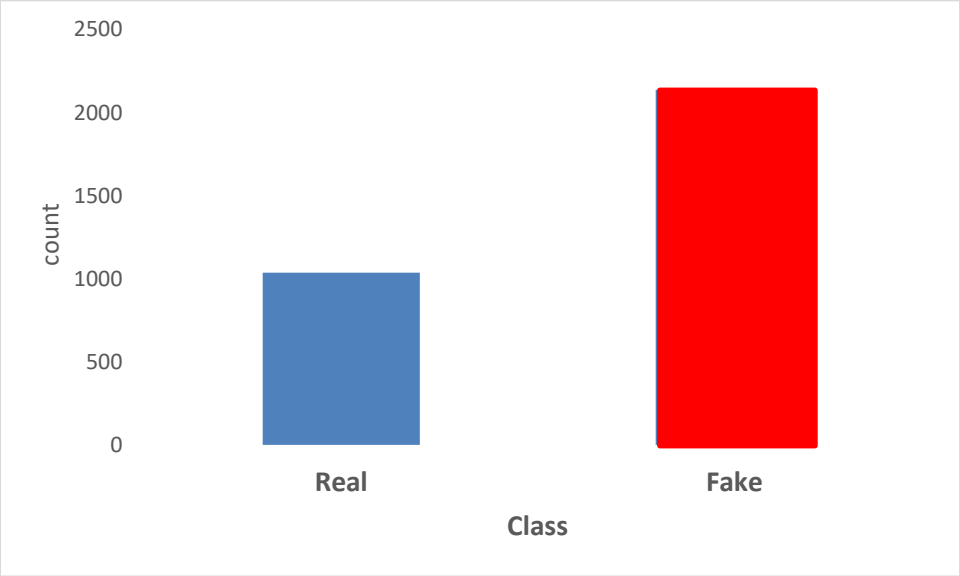


Figure 6: A count plot of the dataset which contains 1037 URL Website and 2137 URL website

```
Epoch 1/50
3900/3900 [==============================] - 1s 233us/step - loss: 0.1210 - acc: 0.9677
Epoch 2/50
3900/3900 [==============================] - 1s 230us/step - loss: 0.1074 - acc: 0.9710
Epoch 3/50
3900/3900 [==============================] - 1s 230us/step - loss: 0.0967 - acc: 0.9738
Epoch 4/50
3900/3900 [==============================] - 1s 233us/step - loss: 0.0881 - acc: 0.9762
Epoch 5/50
3900/3900 [==============================] - 1s 232us/step - loss: 0.0810 - acc: 0.9782
Epoch 6/50
3900/3900 [==============================] - 1s 230us/step - loss: 0.0750 - acc: 0.9787
Epoch 7/50
3900/3900 [==============================] - 1s 233us/step - loss: 0.0697 - acc: 0.9795
Epoch 8/50
3900/3900 [==============================] - 1s 229us/step - loss: 0.0655 - acc: 0.9823
Epoch 9/50
```

Figure7: Training process of the Deep Learning Model which displays the training steps, loss values and accuracy for 1-9 epochs

## 5. RESULT

After successful training of the model, an accuracy of 99.9% was achieved as show in table 1. Deep learning yielded an accuracy of 99.9% than the existing system(SVM) with an accuracy of 98.8%, recall of 88%, precision of 91.66% and F1 score of 87.7%. figure 8: depicts a sample URL website figure 9 depicts the website URL is real and figure 10 depicts the website URL is Real. Table 1 display a comparative analysis between the existing system and proposed system in term of performance.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107


Figure 8: Sample Web Page


Figure 9: Depicts the website URL is real

  

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

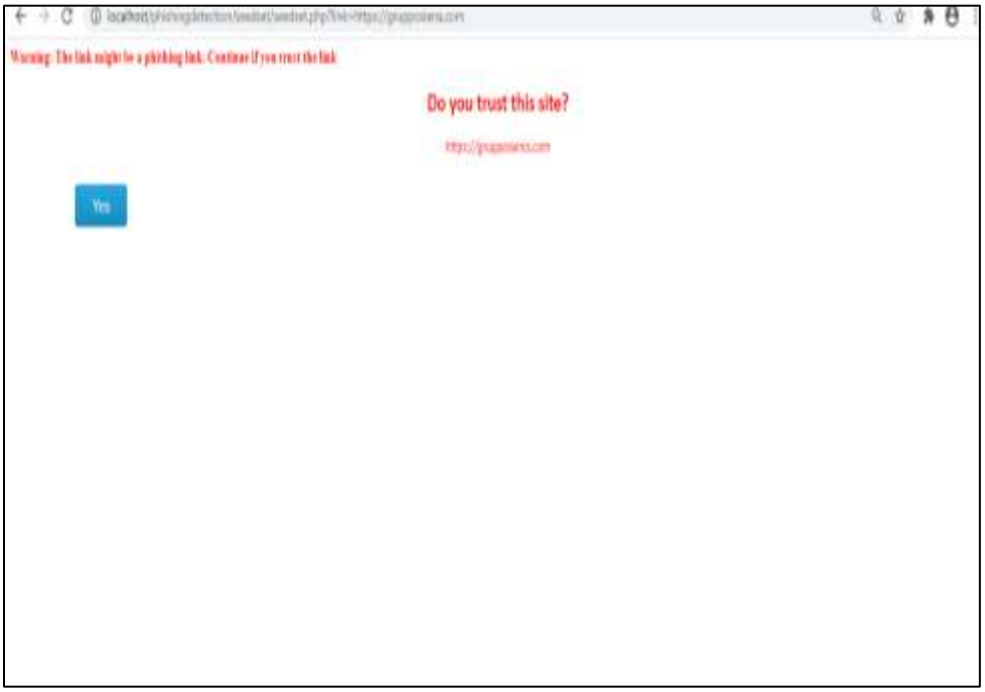


Figure 10: Display that the website URL is Fake

**Comparison with Performance**

The proposed and existing systems have applied Deep learning techniques to the acquisition process to classify the URL website as shown in Table 1. However, the proposed system has used Generated Adversarial Network dataset of 3, 207. While existing system applied Support Vector Machine. The proposed system has produced 98.9%accuracy in the test dataset.

Table 1: Comparison Performance of Proposed System (deep learning) and Existing System (SVM)

| Model | Total URL | Precision | Recall | Accuracy | F1 score |
|---|---|---|---|---|---|
| **Deep learning** | 3207 | 98.7% | 98.1% | 99.9% | 98.4% |
| **SVM Nkue et.al(2021)** [3]. | 3200 | 91.66% | 88% | 98.8% | 89.7% |

To evaluate the effectiveness of our solution, we have used four metrics such as accuracy, precision, recall and F1 score.

$$\text{Testing accuracy} = \frac{(TN+TP)}{(TN+TP+FN+FP}$$

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

$$\text{F1 score} \quad 2TP/(2TP+FP+FN)$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 11, November 2021

DOI: 10.17148/IJARCCE.2021.101107

Where TP is the count of true positive sample, TN is the count of true negative samples, FP is the count of false positive samples and FN is the count of false negative samples from a confusion matrix. The table present the performance matrix for SVM and deep learning. Deep learning scores the highest percentage of 99.9% for accuracy, 98.1% for recall, 98.7% for precision and 98.4% for F1score.

## 6. CONCLUSION AND RECOMMENDATION

This paper presents a Tensorflow and Keras framework in building a Deep Neural Network in classifying URL websites. This Deep Learning Algorithms uses dataset 3,207 URL websites, 1,037 are Real URL websites and 2,137 are fake. The model was built and trained with a total number of 8672 input neurons, 1 output and one dense layer, a batch size (batch size equals the total dataset thus making the iteration and epochs values equivalents) of 20 and epoch (The training steps) value of 30.After successful building of the model, an accuracy of 99.9% was achieved. This paper can further be extended by deploying the trained model into an android application where users can detect a fake URL Website sent to their phone to detect if it a Fake URL or Real URL.

## REFERENCES

[1] Liu, J., & Ye, Y. (2001). Introduction to e-commerce agents: marketplace solutions, security issues, and supply and demand. *In E-commerce agents, marketplace solutions,* security issues, and supply and demand, 1-6.

[2] APWG "Phishing activity trends report 3rd quarter."*US.* Vol 1Issue 11, 2018.

[3] Nkue, D., Daniel, M., & Bennett, E.O(2021) An Efficient Model for Detecting Uniform Resource Locator (URL) Phishing using Machine Learning Techniques *International Journal of Computer Techniques – Volume 8 Issue 3, June 2021 p.46-55.*

[4] Aydin. M. & Baykal, N. (2015). Feature extraction and classification phishing websites based on URL*," IEEE Conference Communication Network Security,* 5,769–770.

[5] Karabatak, M. & Mustafa, T. (2018). Performance comparison of classifiers on reduced phishing website dataset," 6th International Symposium. On *Digital Forensic Security*. *Proceeding,*
1(1), 1–5.

[6] Rishikesh, M. & Irfan,S.(2018)."Phishing Website Detection using Machine Learning Algorithms", *International Journal of Computer Applications,* 23 (181), 0975 – 8887.

[7] Yue Z., Jason, H., lorrie, C. (2007). Cantina: a content-based approach to detecting phishing web sites.

[8] Dogukan, A., Abdullah, A., Ali, A.M.(2017) Detecting Phishing Websites Using Support Vector Machine Algorithm. *2nd World Conference on Technology, Innovation and Entrepreneurship,* Istanbul, Turkey. Edited by Sefer Şener, p.139-142