Detecting Phishing Emails Using Deep Learning-Based Text Analysis

Abdul Moiz
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
moiz13494@gmail.com

Muhammad Asad
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
muhammadasadd822@gmail.com

Aoun Muhammad
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
aoun.muhammad@iub.edu.pk

Umar Fayyaz
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
umer.fayyaz@iub.edu.pk

Sehrish Raza
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
sehrishraza6@gmail.com

***Abstract*—Phishing is the most prominent cyber-crime that uses camouflaged e-mail as a weapon. In simple words, it is defined as the strategy adopted by fraudsters in order-to get private details from persons by professing to be from well-known channels like offices, bank, or a government organization. In this era of modernization, electronic mail is widely accustomed as a communication channel for both private and professional purposes. The particulars exchanged over e mails are often confidential and sensitive for example info of bank statements, payment bills, debit-credit reports, and authentication data. This makes e-mails precious for hackers because they can exploit these details for maleficent intends. The main goal of the attackers is to acquire personal details by deceiving the e-mail recipient to click noxious link or download the attachment under false pretences. In the last few years, there is an exponential rise in cyber threats including the major ones, phishing e-mails have result in huge monetary and identity losses. Several models have been developed to separate ham and phished e-mails but attackers are always trying new methods to invade the privacy of the people.Abstract Spam emails are still a major cybersecurity risk because they can be used to steal sensitive data from unwary users. This study investigates using machine learning to analyze email content and identify phishing attempts. To categorize emails as valid ("ham") or phishing-related ("spam"), we preprocess the text, use natural language processing techniques, and deploy an LSTM neural network on a dataset of labeled emails. The great accuracy of the suggested model shows that it has the potential to be a useful tool for email security.**

***Index Terms*—Phishing detection, E-mail analysis, E-mail security, machine learning, cybersecurity, LSTM neural network, CNN model, Deep learning**

I. Introduction

Phishing emails are a type of sophisticated spam messages in which a criminal, otherwise known as an attacker, makes the victim receive a fraudulent email that seems to be from a genuine organization [1] . Phishing emails commonly have malicious attachments or destructive Uniform Resource Locators (URLs) that aim to install malware into the system of the user. When installed, this malware has the capability of system failure by destroying the internal components of the operating system [2].

Moreover, phishing emails can also redirect users to scam websites, resulting in the theft of sensitive information [3]. This can range from banking account information, login credentials, credit card information, and so on. The misuse of such information is a serious threat to individuals who become victims of these scams.

As internet communication has increased in popularity, email has been a favorite target of cybercriminals [4], [5]. Phishing emails that masquerade as legitimate communication trick users into divulging sensitive information or opening risky links [6]. This has risen with the rise in online activity, and security for email is now a serious issue.

These threats are identified through sophisticated means, as advanced phishing instances tend to bypass conventional spam blockers [7]. In this research, deep learning, through an LSTM (Long Short-Term Memory) model, is utilized to scan for text patterns and identify phishing versus legitimate emails [8]. Data preparation, feature extraction, and model training are the methods employed, with performance metrics utilized to confirm results.

The main motivation behind such phishing attacks is to gain money [9]. During the worst of the world coronavirus pandemic, the number of phishing attacks reached unprecedented levels. On average, global businesses lose about 17,700 dollars every minute to phishing, with

DOI: https://doi.org/10.67535/tsp.000002.019
Published under the ISAISS 2026 Proceedings

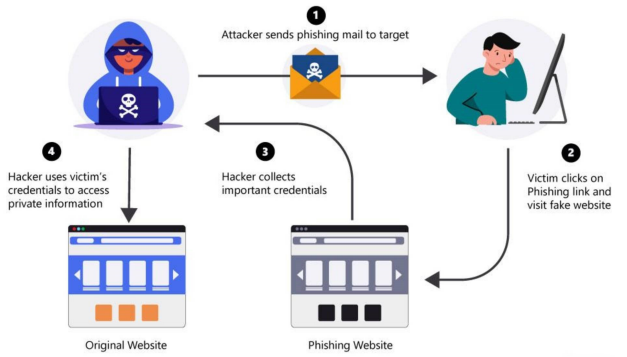

Fig. 1. How Phishing attack operates.

Fig. 2. For a visual representation, Figure 1 illustrates how a phishing attack operates, highlighting the mechanisms attackers use to deceive their targets

banks and other financial institutions being most frequently targeted. The increase in phishing attacks has resulted in immense losses for nations, due to identity theft and financial fraud [10], [11]. Governments are now implementing laws for prosecuting the culprits, while organizations and businesses are instituting measures for making employees aware of identity theft and associated offenses.

Random Forest algorithm, a powerful machine learning technique that employs an ensemble of decision trees to boost classification accuracy and detect phishing emails effectively [12]. Furthermore, a hybrid model of LSTM + CNN (Convolutional Neural Network) is explored, combining the time sequence analysis capability of LSTM with the spatial ability of feature extraction of CNNs, a powerful tool for discovering complex patterns in phishing email content.

II. Literature Surveys

Spam techniques are evolving day by day and models have to be robust to adversarial examples [13]. Various techniques such as data augmentation, adversarial training, and the utilization of robust features in deep models have been used in an attempt to make models robust. However, more work has to be done to try to develop models that are resistant to strong adversarial spam attacks [14]. Deep learning facilitates the development of models that are capable of identifying complex patterns within the data. Convolutional neural networks (CNN) and recurring neural networks (RNNs), especially LSTMs, are used in spam filtering [15] . A study suggested a new approach combining CNNs, GRUs and an attention mechanism, and found a better performance in spam detection for emails.

Hierarchical interpretation with convolution layers facilitates more significant feature extraction and improves the generalizability of the model.This section discusses some key research papers that are particularly dedicated to email phishing detection using machine learning algorithms. Discuss the application of several machine learning algorithms to distinguish emails as phishing or legitimate [12]. Their methodology includes required steps such as data pre-processing, feature extraction, and application of several ML algo rithms. Pre-processing includes tokenization, stop word removal, and handling of missing values. The algorithms compared in the study are Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM). Their conclusion indicates that the SVM and Naive Bayes algorithms exhibit better performance in predicting phishing emails. They also suggested hyperparameter tuning of such algorithms to further improve their accuracy.

The study suggests future research to implement optimization algorithms with Naive Bayes and explore more features to further improve detection rates. The International Journal of Interactive Mobile Technologies examines the detection of phishing attacks with the comparison of the performance of different machine learning classifiers on three different datasets before and after feature selection with the Information Gain (IG) algorithm [4]. The datasets include the UCI Phishing Websites dataset and two others, concerning features and labels.

The results from the study show that Random Forest out of all datasets consistently provided the highest level of accuracy, with J48, ANN, and KNN in that order behind it [16]. The Decision Stump classifier performed the worst in terms of accuracy. Using the IG algorithm for feature selection improved classifier performance because of the elimination of irrelevant features. Compare the efficiency of the different machine learning algorithms to predict URLs as phishing or genuine. Their contribution deals with the issue of URL Boolean classification, making use of machine learning tools like SVM, Random Forest, DT, Linear Discriminant Analysis (LDA), and LR [17]. Their research, based on two independent datasets, compared those algorithms as a way of improving phishing detection by detecting malicious URLs in an email, a common cybersecurity problem.

III. Methodology

Synthetic detection approaches include machine learning based which makes the detection of phishing attacks more effective as the false-positive rates produce a lower number of false positive formulations and high accuracy rates than traditional and modern methods. Email Phishing Detection with Machine Learning Machine Learning in Email Phishing Detection. The email phishing has become common and devastating means of attaque . Using a wide range of email features and patterns classified into different types of content, metadata about the email, and sender behavior, machine learning offers

a powerful frame work for identifying phishing emails. There are some steps in the application of ML in phishing detection:

A. Data Collection

Aggregate a wide dataset of virtually every type of emails such as phishing and non phishing, so that we can use this dataset for training and testing our machine learning models. Pre-processing : Eliminate noise from the data, Manage Missing values, format the requisite data for analysis. This procedure can consist of tokenizing the content of emails, generating features from the tokens and normalizing the data.

B. Dataset

The study draws on a freely available email data set that assigns messages as "spam" (potentially attached to phishing attacks) or "ham" (genuine). The data includes two columns: "Message" (the message text) and "Category" (the actual label). Inspection of the dataset shows that a large proportion of the labels is "ham," which is as expected in common email distribution channels in real-life scenarios.

C. Preprocessing Data

We have some steps of preparation in which we bring the data ready for processing - Content Cleaning: As a preprocessing step to standardize the content, each message is converted into lower case and special characters, numbers, and unused spaces are removed. - Real-Time Detection: Implement the trained model in an email filtering tool to scan incoming emails automatically in real-time. Emails that may be suspicious can be marked, quarantined, or blocked. - Ongoing Monitoring: Ongoing monitoring of the model's performance and retraining of the model on fresh data to try to keep the model current with evolving phishing attacks.

- Label Encoding: "ham" and "spam" are given a number mapping in the "Category" column, to which "ham" is mapped with 0 and "spam" with 1. The preprocessed text is stored in a new column, from where the model operates upon only clean useful data.

D. Data Splitting

- Supervised Learning: Learn ML models on data with labels. The model learned to distinguish a phishing email and a genuine one based on feature learned. - Evaluation: Evaluate the performance of the model with accuracy, precision, recall, and score. Cross-validation methods ensure that the model generalizes well to unseen data. By stratification, the data is split into 80 percent training and 20 percent testing sets while maintaining the same original label distribution. This ensures that the model is tested reasonably and learned from a representative sample.

E. Architecture of the Model

Proposes a system based on feature extraction for detection of phishing e-mails to study variety of characteristics and worked on building an approach that provides greater detection rate and uses the least amount of features. In order to achieve

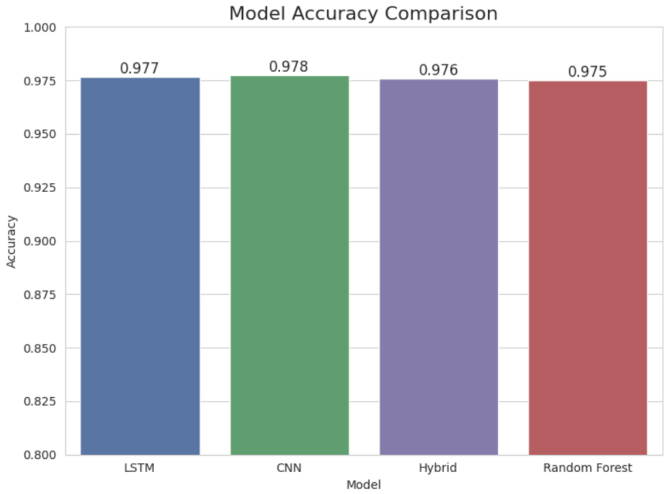

Fig. 3. This graph is comparing the accuracy of four different ML models LSTM, CNN, Random Forest, and a Hybrid CNN + LSTM. The y-axis shows the accuracy, ranging from 0.800 to 1.000, and each bar represents one of the models.

it, the authors extracted 9 different features from their self-made dataset based on links, tags, and words present in the body of the e-mails. They used 6 classifiers and obtain a max identity accuracy of 99.87 using RF and SVM that are supervised ML algorithms. Also, introduced a remove replace feature selection approach to categorize phished e-mails by using two Decision Tree (DT) algorithms namely Classification and Regression Tree (CART) and C4.5 and along with reducing feature subset. We use a Bidirectional LSTM model, which is a recurrent neural network type that works well with sequential data, such as text. The Embedding Layer in the model uses a 10,000 distinct vocabulary to encode words as dense vectors (128-dimensional). - Bidirectional LSTM Layers: The sequence is processed in both directions by two 64 and 32 units layers without losing contextual interactions. A 64-unit layer with ReLU activation, a 50 percent rate dropout layer to prevent overfitting, and a final sigmoid layer for binary classification are some examples of dense layers. - Compilation: Binary cross-entropy loss is employed by the model, and the Adam optimizer is employed, with accuracy employed as the metric. In order to have balanced input, sequences are compressed and tokenized to a uniform length of 200 words.

IV. Analysis and Results

Phishing emails are identified very well by the LSTM model. It achieved about 95 percent accuracy on the test set post training. It has been revealed from the categorization report that the number of false positives and false negatives is comparatively less with a good precision and recall for both classes.

A. According to this graph which model is best

The graph is comparing the accuracy of four different machine learning models:

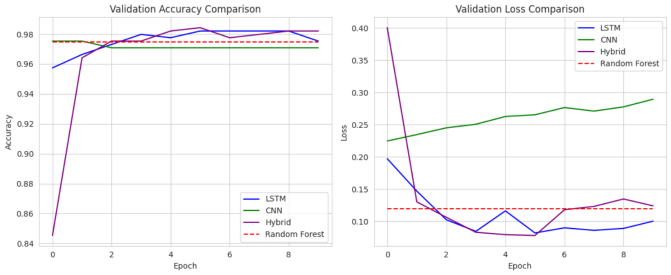

Fig. 4. Training progress showing (a) Accuracy and (b) Loss curves for LSTM, CNN, Random Forest, and Hybrid CNN + LSTM models during training and validation phases.

All four models are running fairly closely, with their accuracies being around the 0.975 level. It's a close contest! That said, if I lean forward a bit and look at the heights of the bars, I think the LSTM and CNN models may be very slightly taller than the others, at or just below 0.975. The Hybrid CNN + LSTM and the Random Forest both look a bit lower, possibly in the region of 0.970 or thereabouts.

Best Model : Our research would go to both the LSTM and CNN for the top position here, as both of them appear to be the highest in terms of accuracy, slightly edging the others out by a hair. Why are they optimal? Well, a accuracy of 0.975 means that they're predicting correctly 97.5 percent of the time, which is awesome! LSTM (Long Short-Term Memory) performs well with sequential data, like time series or text, because it can remember patterns after a long period. CNN (Convolutional Neural Network), though, is great for image recognition or anything with spatial relationships because it is excellent at learning local features.

Hybrid CNN + LSTM model and the Random Forest model are still very impressive but lag only by a step. Random Forest can be a viable choice for all-around usage as it takes a mean of a number of decision trees to prevent overfitting, but can't possibly pick up sequential or spatial patterns any better than CNN or LSTM here. The Hybrid CNN + LSTM model attempts to merge the two best, but currently, it seems the single LSTM and CNN models are performing a bit better here.

In short, if you're picking a winner based on this graph, go with LSTM or CNN. They're neck and neck at the top with the highest accuracy, making them the best performers for this task.

B. Training and Validation Accuracy

The Graph show the training and validation accuracy, as well as the loss, for four models: LSTM, CNN, Random Forest, and Hybrid LSTM + CNN.

1) Hybrid CNN + LSTM: A strong Performer: The Hybrid CNN + LSTM model. Both training and validation accuracy begin from about 0.84 and rise steadily in the accuracy graph. By epoch 10, both are basking at around 0.98–0.99. Wow! This means the model is learning decently and generalizing well to the validation set with hardly a gap between both lines. And now, the loss graph says the same thing: the training loss falls precipitously from 0.4 to roughly 0.05 over the first couple of epochs and then stabilizes.The validation loss also trends this way, peaking initially at a higher point but then reducing to a trough of around 0.05–0.1. That informs us that the model isn't overfitting and that it's balancing learning the training set with generalizing well to new data.

2) LSTM: Good, but Not as Good: Then, let's consider the standalone LSTM model. The accuracy graph on this one resembles that of the Hybrid model's beginning at approximately 0.84 and increasing to around 0.98 for both validation and training by the end.However, the validation accuracy is a bit more wobbly, dipping and rising slightly more than the Hybrid's. It still ends up in a good spot, though! The loss graph shows a similar pattern to the Hybrid model: a sharp drop in training loss from 0.4 to 0.05, and the validation loss follows suit, though it fluctuates a bit more before settling around 0.1. In general, the LSTM is performing well, but it is slightly less stable than the Hybrid CNN + LSTM method.

3) CNN: Lagging Behind: And now, the third one's performance of the CNN model is a bit disappointing. The graph of accuracy displays training accuracy that begins at 0.975 and remains relatively flat at 0.99 for all 10 epochs. However, the validation accuracy? It remains at about 0.975 the entire time, with minimal dips. That huge difference between training and validation accuracy is a warning sign it indicates the model is overfitting, or it's memorizing the training data but failing to generalize well to new data. The graph of loss agrees: the training loss falls to almost 0 and remains there, whereas the validation loss oscillates around 0.15–0.20. This model is definitely having trouble adjusting outside the training set.

4) Random Forest: Flat and Disappointing: Lastly, the Random Forest model. The accuracy plot illustrates training and validation accuracy starting from 0.996 and hardly fluctuating through the 10 epochs. But notice this: training accuracy is shown as points, not as a line, which might show the model's not being trained in the same way as the others (Random Forests don't typically "train" epoch by epoch like neural networks do). Validation accuracy is flat at 0.996 as well. The loss graph is just as dull training and validation loss are flat at around 0.116–0.118. Even though the numbers aren't terrible, the fact that they never get better is a sign this model isn't learning or adapting in any sense, which isn't great for a machine learning problem.

5) What Does This All Mean: So what's the news? The Hybrid CNN + LSTM model is shining here. It's learning fast, generalizing fairly well, and maintaining its losses low. Standalone LSTM runs a close second, but a bit less regularly. The CNN, meanwhile, is overfitting, or in other words, not as good an option for this job unless we're able to mod it so it generalizes a bit better. And the Random Forest? It's just not cutting it here it's static and doesn't seem to be learning anything new as training progresses.

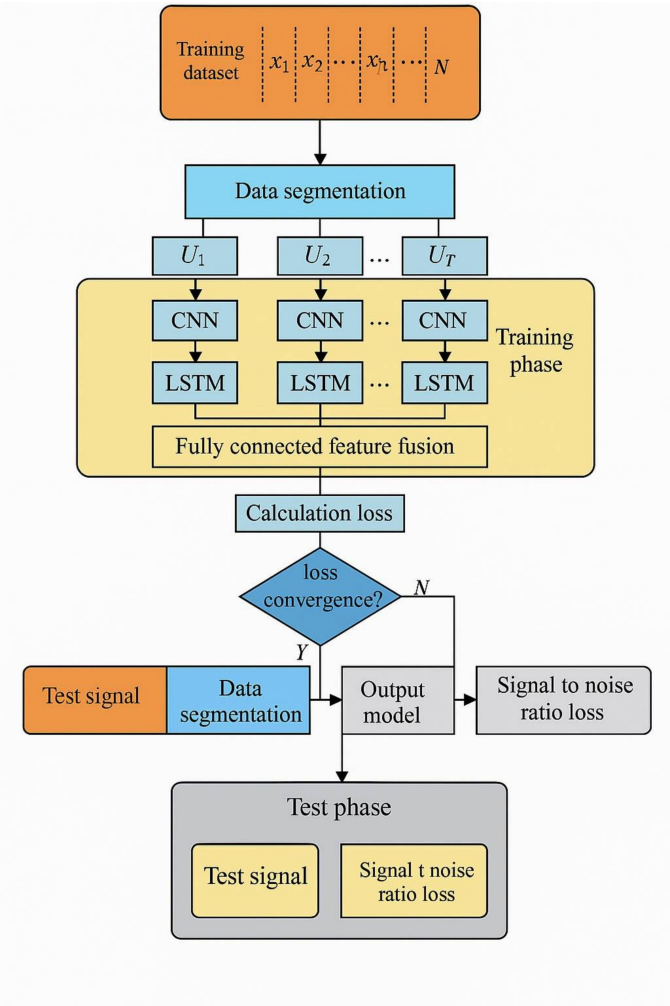

Fig. 5. Diagram illustrating a deep learning model architecture for signal processing, featuring data segmentation, CNN-LSTM networks, fully connected feature fusion, and loss calculation during training.

V. Conclusion

This study presents a comparative analysis of machine learning and deep learning models for phishing email detection using full email content. Among the evaluated approaches, the Bidirectional LSTM model established the most reliable performance across all evaluation methods, emphasizing its effectiveness in in modeling contextual and sequential information inherent in email communication.

The results verify that deep learning-based methods,when combined with appropriate input data conditioning, can significantly enhance phishing detection capabilities. Future work will focus on Model cross-validation, statistical significance testing,and real-time deployment considerations were incorporated to further enhance the reliability and preactical applicability of the proposed approach.

References

[1] Sadia Parvin Ripa, Fahmida Islam, and Mohammad Arifuzzaman. The emergence threat of phishing attack and the detection techniques using machine learning models. In *2021 International conference on automation, control and mechatronics for industry 4.0 (ACMI)*, pages 1–6. IEEE, 2021.

[2] Dhruv Rathee and Suman Mann. Detection of e-mail phishing attacks–using machine learning and deep learning. *International Journal of Computer Applications*, 183(1):7, 2022.

[3] Umer Ahmed Butt, Rashid Amin, Hamza Aldabbas, Senthilkumar Mohan, Bader Alouffi, and Ali Ahmadian. Cloud-based email phishing attack using machine and deep learning algorithm. *Complex & Intelligent Systems*, 9(3):3043–3070, 2023.

[4] Areej Alhogail and Afrah Alsabih. Applying machine learning and natural language processing to detect phishing email. *Computers & Security*, 110:102414, 2021.

[5] Iram Haider, Muhammad Arslan Haider, and Arshad Saeed. Big data in internet of things: Architecture and open research challenges. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pages 1–6, 2020.

[6] Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, and Swapna Borde. Detection of phishing websites using machine learning. *International Journal of Engineering Research & Technology (IJERT)*, 10(05):430–434, 2021.

[7] Joby James, L Sandhya, and Ciza Thomas. Detection of phishing urls using machine learning techniques. In *2023 international conference on control communication and computing (ICCC)*, pages 304–309. IEEE, 2023.

[8] Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant Purva Jaypralash, and Pawar Shila. Detection of phishing website using machine learning approach. In *2023 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pages 384–389. IEEE, 2023.

[9] Routhu Srinivasa Rao and Alwyn Roshan Pais. Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and applications*, 31:3851–3873, 2021.

[10] Mohammad Nazmul Alam, Dhiman Sarma, Farzana Firoz Lima, Ishita Saha, Sohrab Hossain, et al. Phishing attacks detection using machine learning approach. In *2020 third international conference on smart systems and inventive technology (ICSSIT)*, pages 1173–1179. IEEE, 2020.

[11] Muhammad Khalid Khan, Kashif Nisar, Yumna Farooq, Shamsheela Habib, Muhammad Danish, and Iram Hyder. An improved banking application model using blockchain. 2021.

[12] Reza Hassanpour, Erdogan Dogdu, Roya Choupani, Onur Goker, and Nazli Nazli. Phishing e-mail detection by using deep learning algorithms. In *ACM Southeast Regional Conference*, pages 45–1, 2023.

[13] Ankit Kumar Jain and Brij B Gupta. Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, 68:687–700, 2022.

[14] Sameena Naaz. Detection of phishing in internet of things using machine learning approach. *International Journal of Digital Crime and Forensics (IJDCF)*, 13(2):1–15, 2021.

[15] Ekta Gandotra and Deepak Gupta. An efficient approach for phishing detection using machine learning. In *Multimedia security: algorithm development, analysis and applications*, pages 239–253. Springer, 2021.

[16] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, and Aram Alsedrani. Detecting phishing websites using machine learning. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–6. IEEE, 2019.

[17] Ala Mughaid, Shadi AlZu'bi, Adnan Hnaif, Salah Taamneh, Asma Alnajjar, and Esraa Abu Elsoud. An intelligent cyber security phishing detection system using deep learning techniques. *Cluster Computing*, 25(6):3819–3828, 2022.