EUROPEAN JOURNAL OF APPLIED SCIENCE, ENGINEERING AND TECHNOLOGY (ISSN 2786-9342)
VOLUME 3, ISSUE 2 (MARCH-APRIL 2025)
WWW.EJASET.COM

Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering

Purna Chandra Rao Chinta ✉
Sr Technical Support Engineer, Microsoft, USA

Chethan Sriharsha Moore
Sr Technical Support Engineer, Microsoft, USA

Laxmana Murthy Karaka
Software Engineer, Code Ace Solutions Inc, USA

Manikanth Sakuru
Lead Software Engineer, JP Morgan Chase, USA

Varun Bodepudi
Senior Solution Specialist, Deloitte Consulting LLP, USA

Srinivasa Rao Maka
Software Engineer, North Star Group Inc, USA

Article History:
Received: 23.01.2025 Revised: 23.02.2025 Accepted: 26.02.2025 Published: 01.03.2025

ABSTRACT:

The prevalence of cybercrime is directly proportional to the growth in the number of people using the internet. There has been evidence of phishing's extensive usage since its beginning, and it is now the most successful cyberattack vector. According to our findings, phishing is the most prevalent kind of cyberattack, and it employs several techniques to deceive its targets. Phishing attacks using malicious URLs, emails, and websites are rather common. Phishing emails continue to pose significant cybersecurity threats, necessitating robust and intelligent detection mechanisms. Using a large-scale phishing email dataset, this research investigates the creation and assessment of sophisticated ML models for detecting phishing emails. Several ML models were used, including CNN, XGBoost, RNN, and SVM. The best answer was suggested by using the BERT-LSTM hybrid model. Featuring an F1-score 99.24, a recall 99.55%, a precision 99.61%, and an accuracy 99.55%, the BERT-LSTM model accomplished remarkable results. Comparative analysis against existing models, including Naïve Bayes, RNN, and SVM, highlighted BERT-LSTM's superior efficacy in detecting phishing emails. Furthermore, training and testing evaluations demonstrated minimal overfitting and consistent generalisation. This study underscores the potential of BERT-LSTM in real-time phishing email detection systems, offering a reliable solution to mitigate phishing threats effectively.

Keywords: Phishing detection, attacks, ransomware, Phishing e-mail, Machine learning.

Suggested Citation: P.C.R. Chinta, C.S. Moore, L.M. Karaka, M. Sakuru, V. Bodepudi, S.R. Maka, "Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering," *Eur. J. Appl. Sc. Eng. Technol.*, vol. 3(2), pp. 41-54, Mar-Apr 2025. DOI: 10.59324/ejaset.2025.3(2).04

INTRODUCTION

The ability to effectively communicate is crucial in today's society. People often utilise email as a means of more rapid and effective communication. Email is become an unavoidable aspect of daily life. The communication process is now simpler, quicker, and less expensive, thanks to email. The popularity of it has grown [1]. Forrester Research shows that 20 per cent of consumers refuse to open emails or attachments, even if their email looks legitimate, due to their loss of trust. Phishing emails are a severe issue for which there is yet no ideal remedy. No model is flawless, even if there are several nice ones available today. Enhancing phishing detection models is essential because phishing emails may be quite deceptive[2]. The advent of file-encrypting ransomware has made phishing attempts far more deadly, yet these misleading emails still manage to inflict millions of dollars in harm.

The issue of phishing emails has grown in recent years. According to experts, " phishing assaults persist in endangering not only the economy and global security but also companies and consumers"[3]. This highlights the need for robust defences against phishing attempts that originate in email messages to protect vital services like banking.

The importance of phishing email detection to user security has lately garnered considerable attention. As a result, several techniques have been developed to detect phishing emails. These methods range from content-based screening to communication-oriented approaches, including authentication protocols, white-listing, and blacklisting [4]. Blacklisting and whitelisting are not widely employed since they have not been shown to be effective enough in other areas. At the same time, content-based phishing filters are quite effective and have seen extensive application[5]. To combat this, researchers have been focussing on content-based procedures, such as developing ML and data mining algorithms that leverage the contents and headers of emails [6].

ML is a branch of AI concerned with the problem of automatically teaching computers new tasks. Supervised ML methods are used to handle classification in our model [7]. In supervised learning systems, the kind of unknown data is predicted by comparing it to known events. ML includes algorithms such as these that continuously learn from data.

Significance and Contribution

The detection of phishing emails is a critical challenge in modern digital communication, as phishing attacks significantly threaten users' security, privacy, and financial assets. This study holds immense significance in safeguarding individuals and organisations from sophisticated phishing threats that exploit weaknesses in email security. The contribution of this study is reflected in its systematic and comprehensive evaluation of ML and DL models to detect phishing emails using a phishing email Dataset. The primary contributions include:

- Collect the phishing email dataset, combining phishing and legitimate email data, to establish a reliable foundation for classification.
- Designed an effective data preprocessing, including tokenisation, removal of stop words, punctuation, and irrelevant features, for improved model performance.
- Implemented advanced ML models, like CNN, XGBoost, BERT-LSTM, and SVM, to classify phishing and non-phishing e-mails.
- Conducted detailed comparisons of models using metrics like F1-score, precision, recall, and accuracy to identify the most effective detection model.

Structure of the Paper

The study is structured as follows: Section II presents relevant work on phishing detection. Section III details the procedures and materials used. Section IV presents the experimental findings of the proposed system. The inquiry is summarised and concluded in Section V.

LITERATURE REVIEW

This section discusses the surveys and reviews articles on Phishing Email Detection System Using Machine Learning.

In This study, Niu et al. (2018) proposes the CS-SVM model. 23 characteristics are retrieved by the CS-SVM in order to construct the hybrid classifier. By combining SVM with Cuckoo Search (CS), the hybrid classifier optimises the RBF parameter selection. A total of 20,071 emails that are not phishing and 1,384 are used as experimental data. An experimental outcome shows that a proposed approach works better in identifying phishing emails than the SVM classifier with the default parameter value. At its best, the CS-SVM classifier attains 99.52% accuracy[8].

In this research, Smadi, Aslam and Zhang, (2018) a novel method is presented that uses reinforcement learning and NNs to detect phishing attempts in an online environment. This is the first time it has happened. The suggested model has the potential to strengthen the system via reinforcement learning, which might lead to a new phishing email detection system that takes into account changes in behaviours that have been recently studied. The suggested method circumvents the issue of a limited dataset by continuously updating the offline dataset with new emails sent in the online mode. They propose a novel approach to sift through the updated information for signs of phishing. They test the proposed technique extensively on widely used datasets and demonstrate that it performs well against zero-day phishing attacks. The results show that the strategy achieves high accuracy at 98.63%, TPR at 99.07%, and TNR at 98.19%. The FPR is 1.81%, and the FNR is 0.93%, both of which are poor. When tested on the same dataset with competing approaches, the suggested model proves to be superior [9].

In this study, Egozi and Verma (2018) Phishing email study doesn't take stop words into account or eliminates them altogether, and punctuation elements are only used in a minimal capacity by the detector. A linear kernel SVM ensemble learning model achieved an 83% TPR and a 96% TNR, regardless of the unusual weighting of parameters such as word counts, stopwords, punctuation, and uniqueness factors. Even when dealing with email data that is noisy, these traits may still be reliably discovered[10].

In this study, Patil, Rane and Bhalekar et al. (2017) put out a system that detects spam emails more accurately by combining the SVM method with the map-reduce paradigm. They also attempt to circumvent the two problems with the SVM by employing the map-reduce approach[11].

In this study, Abutair and Belghith (2017) present the CBR-PDS. Its primary component is the CBR approach. Unlike previous classifiers that need extensive pre-training, the suggested method is very dynamic and flexible, since it can quickly adjust to recognise novel phishing assaults using a comparatively small data set. They run several scenarios on a balanced set of 572 real and malicious URLs to test our system. Experimental findings show that CBR-PDS system achieves accuracy level greater than 95.62% even with limited data sets and tiny collections of features [12].

In this study, Şentürk, Yerli and Soğukpnar et.al. (2017) the use of data mining and ML methods is suggested as a means of phishing detection. Combating phishing attempts via email has a success rate of 89 [6].

Table I summarises various phishing email detection methodologies, datasets, and findings, highlighting approaches. It also outlines limitations and future work, emphasising the need for larger datasets, diverse scenarios, and enhanced model robustness.

Table 1. Summary of Background Study for Phishing Email Detection System Using Machine Learning

Author	Methodology	Dataset	Findings	Limitations & Future Work
Niu et al.	Hybrid Cuckoo Search SVM (CS-SVM) with RBF optimisation	1,384 phishing and 20,071 non-phishing emails	Achieved 99.52% accuracy by optimising the SVM with the Cuckoo Search method and extracting 23 features.	Limited testing on larger datasets; Future work could adapt hybrid approaches for real-time phishing email detection in dynamic environments.

Smadi, Aslam, and Zhang,	Neural network combined with reinforcement learning for online phishing detection.	Various well-known phishing datasets.	High accuracy (98.63%), TPR (99.07%), and TNR (98.19%). Low FPR (1.81%) and FNR (0.93%). Can handle zero-day phishing attacks dynamically.	Handling of evolving phishing techniques over time, and adaptation to new phishing behaviors needs further refinement.
Egozi and Verma et.al.	Ensemble learning with linear kernel SVM and unconventional features	Spam email data	Unconventional characteristics yielded an 83% TPR and a 96% TNR; robustness was maintained even in the presence of noisy email data.	Limited to specific features like stop words and punctuation; Future work could refine feature engineering and test additional models for increased robustness.
Patil, Rane and Bhalekar et.al.	SVM integrated with Map-Reduce paradigm	Phishing data	Improved accuracy and scalability in spam email detection using the Map-Reduce paradigm to address SVM limitations.	Focused on spam emails instead of phishing; Future work could extend this approach to detect phishing with large-scale distributed systems and other ML classifiers.
Abutair and Belghith	Case-Based Reasoning (CBR) phishing detection system.	Balanced 572 phishing and legitimate URLs.	Despite having a tiny set of features and insufficient data, the CBR-PDS system managed to attain an accuracy of 95.62%.	Further testing needed on more diverse phishing attack scenarios and larger datasets.
Şentürk, Yerli and Soğukpnar et.al.	ML and data mining techniques	Email messages (dataset specifics not mentioned)	Achieved an 89% success rate against phishing attacks from email messages.	Relatively low success rate compared to newer methods; Future work can focus on improving feature extraction and applying ensemble models for better performance.

METHODOLOGY

In this methodology, order to be able to classify the phishing emails will affect the efficiency of a model in detecting the phishing emails. The following implementation steps begin with data collection (phishing Email dataset). Data preprocessing includes tokenisation, stemming, and removal of stop words, numbers, and punctuation, along with phishing term weighting and feature extraction techniques. A dataset is divided into training (70%) and testing (30%) sets. A training set is utilised to train several ML models, such as CNN, XGBoost, RNN, and SVM. The models are then evaluated on unseen data using metrics like recall, accuracy, precision, and F1-measure. After evaluating the models' performance, the top-performing one is included in an intelligent phishing email detection system that operates in real-time to ensure accurate categorisation. The workflow of phishing email detection on the phishing email dataset is shown in Figure 1:

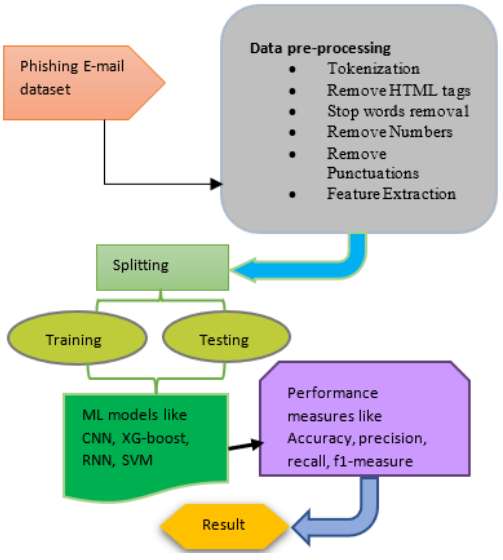

Figure 1. Workflow of Phishing E-mail Detection

Overall implementation steps for phishing email detection are explained below:

Data Collection

A dataset of email messages curated or generated with the express purpose of studying and analysing phishing attempts is called an email phishing dataset. Typical components of such datasets are a collection of phishing emails collected from diverse sources, such Spam Assassin and the UCI ML library. The Distribution of phishing email histogram is provided in Figure 2.

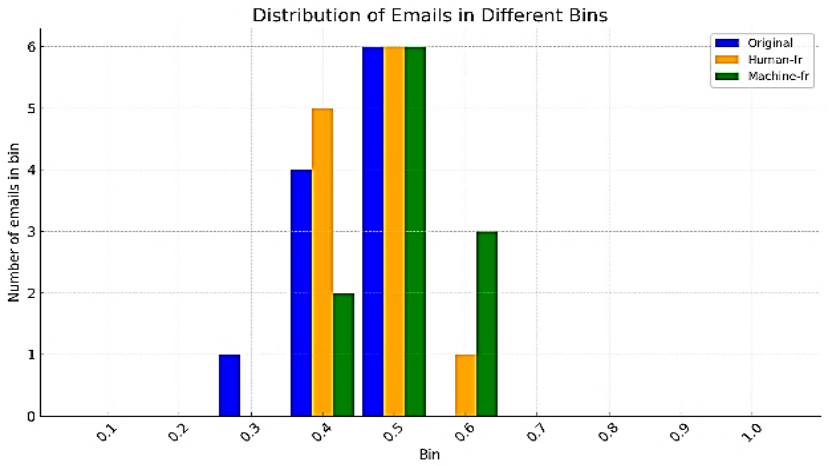

Figure 2. Histogram for Phishing Emails

Figure 2 displays interleaved histograms for both the original emails and the two revisions. Separately, each figure displays the outcomes for a specific email category. The purpose of this presentation is to demonstrate the overall shift in "demographics" that resulted from each change, as well as the distinct ways in which each change impacted the distribution of the initial email categories across the evaluation space.

Data Preprocessing

An important part of information discovery is getting the data ready for analysis. A number of processes are involved, including data reduction and transformation. The effectiveness and precision of learning algorithms are jeopardised if raw input is transformed into low-quality data. Further processing key terms are as follows:

- **Tokenization:** The email's topic and body text are analysed and used to produce tokens [13][14]. The retrieved tokens are all normalised such that any tokens with morphological or flexional ends are deleted.

- **HTML tags removal:** Emails may have their HTML tags stripped using a parser or regular expressions; these tags are used to organise and style online content.

- **Remove Punctuations:** Regular expressions and string manipulation methods may be used to remove punctuation marks and special characters from the dataset. As a result, the model may handle the input more efficiently, and the vocabulary size may be reduced.

- **Remove Stop words**: "Stop words" are overused terms that detract from NLP for deep learning [15]. The words "a", "an", "the", "and", "but", and so on are all stopped words. It is possible to filter out stop words from a dataset by using a stop word filter that is part of an NLP package or by utilising a list of prohibited words.

- **Remove Numbers:** It is possible to filter out non-relevant numerical data from the dataset in order to improve phishing email detection [14]. This has the potential to reduce background noise and enhance the model's pattern-learning capabilities.

Feature Extraction

An essential part of text classification is feature extraction, which involves transforming text input into a numerical representation that DL models can understand [16]. The primary goal of this stage is to get useful textual data for the classifier's training and evaluation [17]. Tokenising the whole text is a common way to do feature extraction. In order to do tokenisation, the text is first deconstructed into its component words and phrases.

Data Splitting

One set of data was utilised for testing, while the other was utilised for training. A data is divided as follows: 80% for training and 20% for testing.

Classification of BERT-LSTM

This section delves into the categorisation of BERT and LSTM. LSTM is an RNN variation that addresses a major issue with standard RNNs: vanishing gradients. LSTM models are better able to deal with long-term dependencies in sequential data because they employ a unique kind of memory cell that can selectively remember or forget information by earlier timesteps [18]. The three basic gate structures of LSTM neurones allow them to choose passed information. Sigmoid neural layers and point-by-point multiplication are the key components of gate construction[19]. An integer between zero and one, representing the degree of forgetfulness, may be produced by the forgetting gate after processing the input. All the data is considered "remembered" if the result is 1. By definition, all data is "forgotten" if it equals 0.

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f \quad (1)$$

In Equation (1), Wf represents the forgetting gate's weights, [h t – 1, xt] combines two vectors together, bf is the thinking bias of the memory gate, σ represents sigmoid function.

The input gate layer is responsible for identifying the values that need updating. Input gate and tanh layers collaborate to update the state, and tanh also creates a new vector.

$$i_t = \sigma(W_t . [h_{t-1}, x_t] + b_t \quad (2)$$

In Equation (2), Wi denotes an input gate's weights, bi denotes an input gate'sbias.

NLP tasks include language interpretation, translation, and text categorisation; BERT is a DL model trained for these

and other similar tasks [20]. The BERT model's foundational transformer design employs self-attention techniques, allowing it to choose concentrate on subsets of the input data in response to specific tasks. BERT learns broad language representations using pre-training, which involves training the model on a large corpus of textual data. After the model has been pre-trained, a few task-specific layers may be added to it to improve BERT's performance on a particular job, such text categorisation[21]. The pre-trained BERT model is fine-tuned using a tagged dataset of text data before it is utilised for text categorisation. After the BERT model has been fine-tuned, it may be applied to pre-trained and unknown text data to categorise it using the representations it has learnt.

The given dataset was used to train and evaluate the BERT with a hybrid classifier, which was built in Python.

In Equations (3-5), respectively. Learning is used to generate the three matrices, namely WQ, WK, and WV. The question is represented by the qi, the key is by the ki, and the information about the token is represented by the vi [22]. To demonstrate self-attention, they use token x1.

$$x_t.W^Q = qt \tag{3}$$

$$x_t.W^K = k_t \tag{4}$$

$$x_t.W^V = v_t \tag{5}$$

In Equations (6)–(8), respectively. WQ, WK, and WV, matrices are each obtained by means of a learning procedure. The question is represented by qi, the key is by ki, and the information about the token is represented by vi. To demonstrate self-attention, they use token x1

$$q_1.k_f = \alpha_{1,f} \tag{6}$$

$$\alpha'_{1,f} = \frac{\exp(\alpha_{1,f})}{\sum_f \exp(\alpha_{1,f})} \tag{7}$$

$$y_1 = \sum_t \alpha'_{1,f} v_t \tag{8}$$

$$self\_attention(Q,K,V) = softmax\left(\frac{QK^t}{\sqrt{dk}}\right)V \tag{9}$$

To summarise, the model uses an attention method to determine contextual connections among the URL's component tokens. Equations (3)–(5) are used to compare each token's attention score to that of other tokens. Lastly, in order to determine whether the URL is malicious, Equation (9) is used to extract the whole semantic meaning of the full URL

Evaluation Metrics

While accuracy measurements are the basis for ML classifier performance evaluations, they use cross-validation as our assessment approach and report on a number of other metrics [23]. Here are four potential outcomes for each email: TP (properly classified phishing email), TN (ham email correctly classified), FP(ham email incorrectly categorised as phishing), and FN(phishing email wrongly classed as ham). Declare accuracy, that is, the percentage of properly categorised emails, for the sake of comparison. It is crucial to provide conventional metrics like recall, accuracy, and F-measure in addition to the FPR and FNR. Their definitions are as follows:

Accuracy

A metric's accuracy may be defined as the percentage of test samples that were properly identified out of all samples in the dataset. Here is the formula for calculating the accuracy metric: (10):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Precision

Partitioning the sum of the positive outcomes that the classifier program accurately predicted yields the precision. This is how it is computed (11):

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Recall

The whole number of valid positive findings is then divided by the sum of all relevant samples. The formula is (12):

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

F1-Measure

Its purpose is to determine how accurate a test is. When the Precision and Recall metrics are added together, the result is the F1 score. The F1-score may take values between zero and one. It provides feedback on the robustness and accuracy of your classifier. This is the formula (13):

$$CapF1 - \text{measure} = 2 * \frac{1}{(\frac{1}{Precision})+(\frac{1}{Recall})} \quad (13)$$

These evaluation parameters predict the performance of models for diabetes disease.

RESULT ANALYSIS AND DISCUSSION

For the experiment implementation, use Python programming language with a Jupyter Notebook environment. An Intel Core i7, 2.2 GHz CPU, "Microsoft Windows 10," and 16 GB of RAM memory are used to run the following models on a personal computer (PC). The proposed model BERT-LSTM is compared (see Table III) with existing models like Naïve-Bayes (NB)[24], Recurrent Neural Network (RNN)[25], SVM[26]. All models were trained on the phishing email datasets, where the BERT-LSTM model outperformance and achieved the highest accuracy provided in Table 2.

Table 2. Results of BERT-LSTM Hybrid Model for Phishing Email Detection

Measures	BERT-LSTM
Accuracy	99.55
Precision	99.61
Recall	99.55
F1-score	99.24

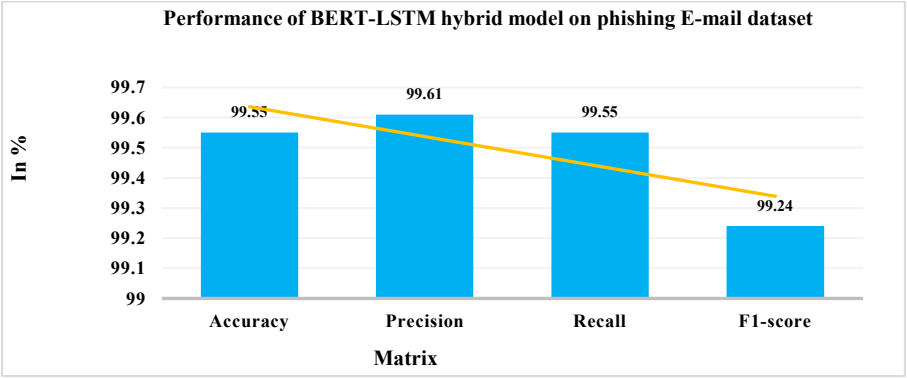

Figure 3. Performance of BERT-LSTM Model

A model's ability to identify phishing emails is seen in Figure 3 and Table II. In this figure, the BERT-LSTM model achieves 99.61%accuracy, precision 99.87%, recall 99.23 &f1-score 99.55%. The model accurately detects phishing e-mails.

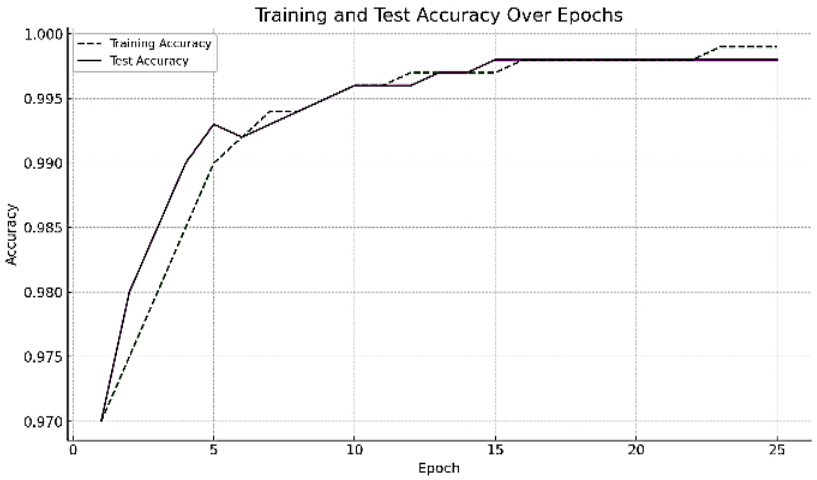

Figure 4. Training and Testing Accuracy for BERT-LSTM

The BERT-LSTM Hybrid model's training and testing accuracy across 1000 epochs are shown in Figure 4. With values between 0.996 and 0.998, the training and validation accuracies are quite excellent. The model's strong performance on both the training and validation datasets implies that it is capable of excellent generalisation and does not exhibit substantial overfitting.

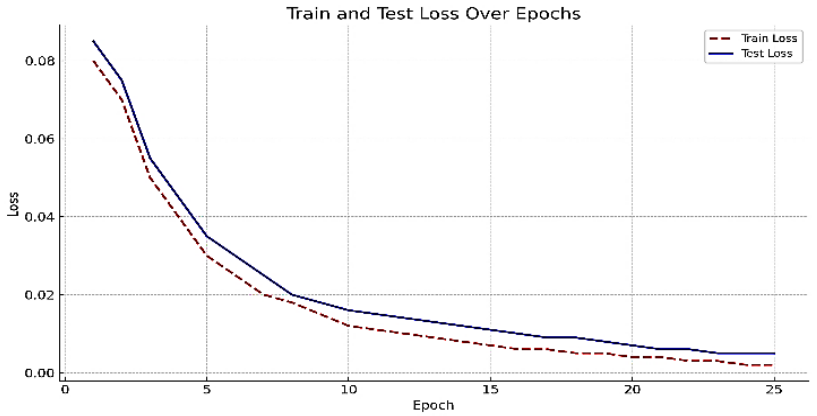

Figure 5. Training and Testing Loss for BERT-LSTM Model

The BERT-LSTM model's training and validation losses across 100 epochs are shown in Figure 5. A training loss, displayed by a red line, reduces more quickly than the test loss, which is likewise red but falls more slowly. These results show that the model is picking up new information and becoming better at what it does.

Table 3. Comparison between ML Models on the Phishing E-mail Dataset

Models	Accuracy	Precision	Recall	F1-score
BERT-LSTM	99.55	99.61	99.55	99.24
NB	70.93	64.70	92.08	76.00
RNN	86	96.2	87.6	91.7
SVM	96.9	72	56.3	63.2

Table 3 compares a performance of ML models like BERT-LSTM, Naïve-bayes, RNN, and SVM across accuracy, precision, recall, and F1-score metrics. BERT-LSTM demonstrates the best performance with an accuracy 99.55%, precision 99.61%, recall 99.55%, and an F1-score 99.24, indicating its exceptional ability to handle the classification task effectively. SVM follows with an accuracy 96.9%, precision 72%, recall 56.3%, and F1-score 63.2, showing balanced and strong performance. The RNN achieves 86% accuracy, with high recall at 96.2% and precision of 87.6%, but its F1-score drops to 91.7, highlighting inconsistency in overall performance. Naïve-Bayes performs the worst, with an accuracy 70.93%, precision 64.70%, recall 92.08%, and an F1-score 76, reflecting significant struggles in identifying positive cases. Overall, BERT-LSTM outperforms the other models.

CONCLUSION AND FUTURE WORK

An email that seems to be authentic but is really created by a phisher to trick the recipient is known as a phishing email. The purpose of phishing emails is to trick recipients into visiting a malicious website that seems authentic. Getting consumers to unknowingly download harmful attachments is another strategy. To assist readers in understanding phishing attempts and emails better, this article provides a brief summary of the topic. With an F1-score 99.24%, an accuracy 99.55%, a precision 99.61%, and a recall 99.55%, the suggested BERT-LSTM hybrid model has shown remarkable performance in identifying phishing emails. Through comprehensive preprocessing and feature extraction, the model effectively captures complex patterns in phishing emails, outperforming traditional models like Naïve Bayes, RNN, and SVM. Training and testing evaluations further reveal its strong generalisation capabilities with minimal overfitting, making it a robust solution for real-time phishing email detection. Despite its success, the model has certain limitations, including a reliance on large labelled datasets for training and significant computational requirements, which could pose challenges for deployment in resource-constrained environments. Additionally, the model may require frequent updates to address rapidly evolving phishing tactics, potentially increasing maintenance efforts.

REFERENCES

[1] M. R. Kishore Mullangi, Vamsi Krishna Yarlagadda, Niravkumar Dhameliya, “Integrating AI and Reciprocal Symmetry in Financial Management: A Pathway to Enhanced Decision-Making,” *Int. J. Reciprocal Symmetry Theor. Phys.*, vol. 5, no. 1, pp. 42–52, 2018.

[2] V. V. Kumar, M. K. Pandey, M. K. Tiwari, and D. Ben-Arieh, “Simultaneous optimisation of parts and operations sequences in SSMS: A chaos embedded Taguchi particle swarm optimization approach,” *J. Intell. Manuf.*, 2010, doi: 10.1007/s10845-008-0175-4.

[3] I. R. Ahamid, J. Abawajy, and T. H. Kim, “Using feature selection and classification scheme for automating phishing email detection,” *Stud. Informatics Control*, 2013, doi: 10.24846/v22i1y201307.

[4] V. V. Kumar, S. R. Yadav, F. W. Liou, and S. N. Balakrishnan, “A digital interface for the part designers and the fixture designers for a reconfigurable assembly system,” *Math. Probl. Eng.*, 2013, doi: 10.1155/2013/943702.

[5] L. M. Form, K. L. Chiew, S. N. Sze, and W. K. Tiong, “Phishing email detection technique by using hybrid features,” in *2015 9th International Conference on IT in Asia: Transforming Big Data into Knowledge, CITA 2015 - Proceedings*, 2015. doi: 10.1109/CITA.2015.7349818.

[6] Ş. Şentürk, E. Yerli, and İ. Soğukpnar, “Email phishing detection and prevention by using data mining techniques,” in *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 2017. doi: 10.1109/UBMK.2017.8093510.

[7] R. Hassanpour, E. Dogdu, R. Choupani, O. Goker, and N. Nazli, “Phishing e-mail detection by using deep learning algorithms,” 2018. doi: 10.1145/3190645.3190719.

[8] W. Niu, X. Zhang, G. Yang, Z. Ma, and Z. Zhuo, “Phishing emails detection using CS-SVM,” in *Proceedings - 15th IEEE International Symposium on Parallel and Distributed Processing with Applications and 16th IEEE International Conference on Ubiquitous Computing and Communications, ISPA/IUCC 2017*, 2018. doi: 10.1109/ISPA/IUCC.2017.00160.

[9] S. Smadi, N. Aslam, and L. Zhang, “Detection of online phishing email using dynamic evolving neural network based on reinforcement learning,” *Decis. Support Syst.*, 2018, doi: 10.1016/j.dss.2018.01.001.

[10] G. Egozi and R. Verma, “Phishing email detection using robust NLP techniques,” in *IEEE International Conference on Data Mining Workshops, ICDMW*, 2018. doi: 10.1109/ICDMW.2018.00009.

[11] P. Patil, R. Rane, and M. Bhalekar, “Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm,” in *Proceedings of the International Conference on Inventive Systems and Control, ICISC 2017*, 2017. doi: 10.1109/ICISC.2017.8068633.

[12] H. Y. A. Abutair and A. Belghith, “Using Case-Based Reasoning for Phishing Detection,” in *Procedia Computer Science*, 2017. doi: 10.1016/j.procs.2017.05.352.

[13] V. V. Kumar, F. T. S. Chan, N. Mishra, and V. Kumar, “Environmental integrated closed loop logistics model: An artificial bee colony approach,” in *SCMIS 2010 - Proceedings of 2010 8th International Conference on Supply Chain Management and Information Systems: Logistics Systems and Engineering*, 2010.

[14] M. Z. Hasan, R. Fink, M. R. Suyambu, and M. K. Baskaran, “Assessment and improvement of intelligent controllers for elevator energy efficiency,” in *IEEE International Conference on Electro Information Technology*, 2012. doi: 10.1109/EIT.2012.6220727.

[15] S. C. R. Vennapusa, T. Fadziso, D. K. Sachani, V. K. Yarlagadda, and S. K. R. Anumandla, “Cryptocurrency-Based Loyalty Programs for Enhanced Customer Engagement,” *Technol. \& Manag. Rev.*, vol. 3, pp. 46–62, 2018.

[16] V. V. Kumar, F. W. Liou, S. N. Balakrishnan, and V. Kumar, “Economical impact of RFID implementation in remanufacturing: a Chaos-based Interactive Artificial Bee Colony approach,” *J. Intell. Manuf.*, 2015, doi: 10.1007/s10845-013-0836-9.

[17] V. K. Yarlagadda and R. Pydipalli, “Secure Programming with SAS: Mitigating Risks and Protecting Data Integrity,” *Eng. Int.*, vol. 6, no. 2, pp. 211–222, Dec. 2018, doi: 10.18034/ei.v6i2.709.

[18] V. V Kumar, M. Tripathi, M. K. Pandey, and M. K. Tiwari, “Physical programming and conjoint analysis-based redundancy allocation in multistate systems: A Taguchi embedded algorithm selection and control (TAS&C) approach,” *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, vol. 223, no. 3, pp. 215–232, Sep. 2009, doi: 10.1243/1748006XJRR210.

[19] M. Z. Hasan, R. Fink, M. R. Suyambu, M. K. Baskaran, D. James, and J. Gamboa, “Performance evaluation of energy efficient intelligent elevator controllers,” in *IEEE International Conference on Electro Information Technology*, 2015. doi: 10.1109/EIT.2015.7293320.

[20] V. Kumar, V. V. Kumar, N. Mishra, F. T. S. Chan, and B. Gnanasekar, “Warranty failure analysis in service supply Chain a multi-agent framework,” in *SCMIS 2010 - Proceedings of 2010 8th International Conference on Supply Chain Management and Information Systems: Logistics Systems and Engineering*, 2010.

[21] V. Kumar and F. T. S. Chan, “A superiority search and optimisation algorithm to solve RFID and an environmental factor embedded closed loop logistics model,” *Int. J. Prod. Res.*, vol. 49, no. 16, 2011, doi: 10.1080/00207543.2010.503201.

[22] V. V Kumar, “An interactive product development model in remanufacturing environment : a chaos-based artificial bee colony approach,” Missouri University of Science and Technology, 2014. [Online]. Available: https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=8243&context=masters_theses

[23] V. V Kumar, M. Tripathi, S. K. Tyagi, S. K. Shukla, and M. K. Tiwari, “An integrated real time

optimization approach (IRTO) for physical programming based redundancy allocation problem," *3rd Int. Conf. Reliab. Saf. Eng.*, pp. 692–704, 2007.

[24] S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of phishing emails using data mining algorithms," in *SKIMA 2015 - 9th International Conference on Software, Knowledge, Information Management and Applications*, 2016. doi: 10.1109/SKIMA.2015.7399985.

[25] R. Vinayakumar, G. H. Barathi, K. M. Anand, K. Soman, and P. Prabaharan, "DeepAnti-PhishNet: Applying deep neural networks for phishing email detection," in *R. Verma, A. Das (eds.): Proceedings of the 1st AntiPhishing Shared Pilot at 4th ACM International Workshop on Security and Privacy Analytics (IWSPA 2018), Tempe, Arizona, USA.*, 2018.

[26] A. Yasin and A. Abuhasan, "An Intelligent Classification Model for Phishing Email Detection," *Int. J. Netw. Secur. Its Appl.*, vol. 8, no. 4, pp. 55–72, 2016, doi: 10.5121/ijnsa.2016.8405.

[27] G. K. Patra, S. K. Rajaram, V. N. Boddapati, C. Kuraku, and H. K. Gollangi, "Advancing Digital Payment Systems: Combining AI, Big Data, and Biometric Authentication for Enhanced Security," *International Journal of Engineering and Computer Science*, vol. 11, no. 08, pp. 25618–25631, 2022, doi: 10.18535/ijecs/v11i08.4698.

[28] S. K. Rajaram, E. P. Galla, G. K. Patra, C. R. Madhavaram, and J. Rao, "AI-Driven Threat Detection: Leveraging Big Data for Advanced Cybersecurity Compliance," *Educational Administration: Theory and Practice*, vol. 28, no. 4, pp. 285–296, 2022, doi: 10.53555/kuey.v28i4.7529.

[29] G. K. Patra, S. K. Rajaram, and V. N. Boddapati, "AI and Big Data in Digital Payments: A Comprehensive Model for Secure Biometric Authentication," *Educational Administration: Theory and Practice*, vol. 25, no. 4, pp. 773–781, 2019, doi: 10.53555/kuey.v25i4.7591.

[30] C. Kuraku, H. K. Gollangi, and J. R. Sunkara, "Biometric Authentication in Digital Payments: Utilizing AI and Big Data for Real-Time Security and Efficiency," *Educational Administration: Theory and Practice*, vol. 26, no. 4, pp. 954–964, 2020, doi: 10.53555/kuey.v26i4.7590.

[31] E. P. Galla et al., "Big Data and AI Innovations in Biometric Authentication for Secure Digital Transactions," *Educational Administration: Theory and Practice*, vol. 27, no. 4, pp. 1228–1236, 2021, doi: 10.53555/kuey.v27i4.7592.

[32] J. R. Sunkara, S. R. Bauskar, C. R. Madhavaram, E. P. Galla, and H. K. Gollangi, "Data-Driven Management: The Impact of Visualization Tools on Business Performance," *International Journal of Management*, vol. 12, no. 3, pp. 1290–1298, 2021.

[33] V. N. Boddapati et al., "Data Migration in the Cloud Database: A Review of Vendor Solutions and Challenges," *International Journal of Computer and Artificial Intelligence*, vol. 3, no. 2, pp. 96–101, Jul. 2022, doi: 10.33545/27076571.2022.v3.i2a.110.

[34] M. S. Reddy, M. Sarisa, S. Konkimalla, S. R. Bauskar, H. K. Gollangi, and E. P. Galla, "Predicting Tomorrow's Ailments: How AI/ML Is Transforming Disease Forecasting," *ESP Journal of Engineering & Technology Advancements*, vol. 1, no. 2, pp. 188–200, 2021.

[35] K. Gollangi, S. R. Bauskar, C. R. Madhavaram, E. P. Galla, J. R. Sunkara, and M. S. Reddy, "Echoes in Pixels: The Intersection of Image Processing and Sound Detection," *International Journal of Development Research*, vol. 10, no. 08, pp. 39735–39743, 2020, doi: 10.37118/ijdr.28839.28.2020.

[36] H. K. Gollangi et al., "Unveiling the Hidden Patterns: AI-Driven Innovations in Image Processing and Acoustic Signal Detection," *Journal of Recent Trends in Computer Science and Engineering*, vol. 8, no. 1, pp. 25–45, 2020, doi: 10.70589/JRTCSE.2020.1.3.

[37] H. K. Gollangi et al., "Exploring AI Algorithms for Cancer Classification and Prediction Using Electronic Health Records," *Journal of Artificial Intelligence and Big Data*, vol. 1, no. 1, pp. 65–74, 2020, doi: 10.31586/jaibd.2020.1109.

[38] S. R. Bauskar et al., "Data Migration in the Cloud Database: A Review of Vendor Solutions and Challenges," SSRN, Jul. 2022, doi: 10.2139/ssrn.4988789.

[39] C. R. M., E. P. G., M. S. R., M. S., V. N. B., and S. K., "Predicting Diabetes Mellitus in Healthcare: A Comparative Analysis of Machine Learning Algorithms on Big Dataset," *Global Journal of Research in Engineering & Computer Sciences*, vol. 1, no. 1, pp. 1–11, 2021, doi: 10.5281/zenodo.14010835.

[40] V. N. Boddapati, E. P. Galla, G. K. Patra, C. R. Madhavaram, and J. R. Sunkara, "AI-Powered Insights: Leveraging Machine Learning and Big Data for Advanced Genomic Research in Healthcare," *Educational Administration: Theory and Practice*, vol. 29, no. 4, pp. 2849–2857, 2023, doi: 10.53555/kuey.v29i4.7531.

[41] G. K. Patra, C. Kuraku, S. Konkimalla, V. N. Boddapati, and M. Sarisa, "Voice Classification in AI: Harnessing Machine Learning for Enhanced Speech Recognition," *Global Research and Development Journals*, vol. 8, no. 12, pp. 19–26, 2023, doi: 10.70179/grdjev09i110003.

[42] J. R. Sunkara, S. R. Bauskar, C. R. Madhavaram, E. P. Galla, and H. K. Gollangi, "Optimizing Cloud Computing Performance with Advanced DBMS Techniques: A Comparative Study," *Journal for ReAttach Therapy and Developmental Diversities*, vol. 6, no. 10s(2), pp. 2493–2502, 2023, doi: 10.53555/jrtdd.v6i10s(2).3206

[43] J. R. Sunkara, S. R. Bauskar, C. R. Madhavaram, E. P. Galla, H. K. Gollangi, M. S. Reddy, and K. Polimetla, "An Evaluation of Medical Image Analysis Using Image Segmentation and Deep Learning Techniques," *Journal of Artificial Intelligence & Cloud Computing*, vol. 2, no. 3, pp. 1–8, Jul. 2023. doi: 10.47363/JAICC/2023(2)388

[44] V. Varadharajan, N. Smith, D. Kalla, F. Samaah, K. Polimetla, and G. R. Kumar, "Stock Closing Price and Trend Prediction with LSTM-RNN," *Journal of Artificial Intelligence and Big Data*, vol. 4, p. 877, 2024. [Online]. Available: https://www.scipublications.com/journal/index.php/jaibd/article/view/877

[45] D. Kalla, D. S. Kuraku, and F. Samaah, "Enhancing Cyber Security by Predicting Malwares Using Supervised Machine Learning Models," *International Journal of Computing and Artificial Intelligence*, vol. 2, no. 2, pp. 55–62, 2021. [Online]. Available: https://www.computersciencejournals.com/ijcai/archives/2021.v2.i2.A.71

[46] G. K. Patra, C. Kuraku, S. Konkimalla, V. N. Boddapati, M. Sarisa, S. K. Rajaram, M. S. Reddy, and K. Polimetla, "Sentiment Analysis of Customer Product Review Based on Machine Learning Techniques in E-Commerce," *Journal of Artificial Intelligence & Cloud Computing*, vol. 2, no. 4, pp. 1–4, Oct. 2023. doi: 10.47363/JAICC/2023(2)389

[47] D. Kalla, N. Smith, F. Samaah, and K. Polimetla, "Hybrid Scalable Researcher Recommendation System Using Azure Data Lake Analytics," *Journal of Data Analysis and Information Processing*, vol. 12, pp. 76–88, 2024.

[48] S. K., G. K. Patra, C. Kuraku, J. R. Sunkara, S. R. Bauskar, M. S. Reddy, and K. Polimetla, "A Comparative Analysis of Network Intrusion Detection Using Different Machine Learning Techniques," *Journal of Contemporary Education Theory & Artificial Intelligence*, vol. 2, no. 1, 2023.

[49] E. P. Galla, H. K. Gollangi, V. N. Boddapati, M. Sarisa, K. Polimetla, M. S. Reddy, and K. Polimetla, "Enhancing Performance of Financial Fraud Detection Through Machine Learning Model," *Journal of Contemporary Education Theory & Artificial Intelligence*, vol. 2, no. 1, 2023.

[50] D. Kalla, N. Smith, and F. Samaah, "Satellite Image Processing Using Azure Databricks and Residual Neural Network," *International Journal of Advanced Trends in Computer Applications*, vol. 9, no. 2, pp. 48–55, 2023.

[51] S. K. Rajaram, S. Konkimalla, M. Sarisa, H. K. Gollangi, C. R. Madhavaram, and M. S. Reddy, "AI/ML-Powered Phishing Detection: Building an Impenetrable Email Security System," *ISAR Journal of Science and Technology*, vol. 1, no. 2, pp. 10–19, 2023.

[52] E. P. Galla, H. K. Gollangi, V. N. Boddapati, M. Sarisa, K. Polimetla, M. S. Reddy, and K. Polimetla, "Prediction of Financial Stock Market Based on Machine Learning Technique," *Journal of Contemporary Education Theory & Artificial Intelligence*, vol. 2, no. 1, 2023. [Online]. Available: https://ssrn.com/abstract=4975786

[53] D. Kalla, N. Smith, and F. Samaah, "Satellite Image Processing Using Azure Databricks and Residual Neural Network," *International Journal of Advanced Trends in Computer Applications*, vol. 9, no. 2, pp. 48–55, 2023.

[54] D. Kalla and N. Smith, "Integrating IoT, AI, and Big Data for Enhanced Operational Efficiency in Smart Factories," *Educational Administration: Theory and Practice*, vol. 30, no. 5, pp. 14235–14245, 2024.

[55] D. S. Kuraku and D. Kalla, "Phishing Website URL's Detection Using NLP and Machine Learning Techniques," *Journal on Artificial Intelligence-Tech Science*, 2023.

[56] D. Kalla, D. S. Kuraku, and F. Samaah, "Enhancing Cyber Security by Predicting Malwares Using Supervised Machine Learning Models," *International Journal of Computing and Artificial Intelligence*, vol. 2, no. 2, pp. 55–62, 2021. [Online]. Available: https://www.computersciencejournals.com/ijcai/archives/2021.v2.i2.A.71

[57] D. Kalla, F. Samaah, S. Kuraku, and N. Smith, "Phishing Detection Implementation Using Databricks and Artificial Intelligence," *International Journal of Computer Applications*, vol. 185, no. 11, pp. 1–11, 2023.

[58] D. Kalla, "Improving E-Commerce Organization Performance Using Big Data Analytics and Artificial Intelligence," Ph.D. dissertation, Colorado Technical University, 2024.

[59] S. Kuraku and D. Kalla, "Emotet Malware—A Banking Credentials Stealer," *IOSR Journal of Computer Engineering*, vol. 22, no. 1, pp. 31–41, 2020.

[60] S. Kuraku, D. Kalla, F. Samaah, and N. Smith, "Cultivating Proactive Cybersecurity Culture Among IT Professionals to Combat Evolving Threats," *International Journal of Electrical, Electronics and Computers*, vol. 8, no. 6, 2023.

[61] D. S. Kuraku and D. Kalla, "Impact of Phishing on Users with Different Online Browsing Hours and Spending Habits," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 12, no. 10, 2023.