Identification And Classification Of Phishing Emails Based on Machine Learning Techniques To Improvise Cyber security

Mani Gopalsamy
Senior Cyber Security Specialist, Louisville, KY, USA- 40220

Abstract- *The incidence of cybercrime is skyrocketing with the proliferation of internet users. Phishing is now the most effective cyberattack vector, with evidence of its widespread use. Typical real-world settings often include an uneven distribution of phishing and benign emails, which causes conventional ML and DL algorithms to incorrectly identify phishing emails and favour benign ones. We discovered that phishing is the most common attack and that many different methods exist to launch one against a victim. The most common phishing attacks include malicious URLs, emails, and websites. The analysis utilises a comprehensive phishing email dataset containing 525,754 instances of phishing and legitimate emails. ML is ushering in a new age for cybercriminals and users concerned about staying protected. Several models were used for phishing email detection, including SVM, DT, BERT-LSTM, and MNB, while performance metrics such as accuracy, precision, recall, and F1-score were measured inside the performance matrix. Results show that the BERT-LSTM model significantly outperforms other models. Specifically, BERT-LSTM achieves the highest accuracy, 99.61%, precision 99.87%, recall 99.23%, and F1-score 99.55%, demonstrating superior capability in detecting phishing emails and minimising false positives.*

Keywords- Phishing Emails, Machine Learning, Detection, Classification, Cybersecurity.

I. INTRODUCTION

Security concerns are growing in number and severity, and the experience of using the Internet has altered dramatically due to the fast evolution of related technology. Cyberattacks against businesses have become much more common due to the widespread use of public resources hosted on the Internet, such as social media, online banking, and cloud computing [1]. In addition to stealing customers' money and personal information, emerging attacks can gravely damage their computers. One major worry among them is phishing, which is an illegal practice that involves using technology and social engineering to steal a victim's identity and account information[2]S. The Anti-Phishing Working Group (APWG) reported that when comparing the first quarter of 2018 to the fourth quarter of 2017, the number of phishing detections rose by 46%. Phishing seems to have been on the rise in recent years, as evidenced by the statistics,just as it is possible to envision damage caused by phishing [3].

Phishing emails: Phishing emails are a kind of cyberattack in which the attacker sends an email to a prospective target, tricking them into thinking it came from a reliable source. The victims are encouraged to do specific things once they gain their confidence, which they would not have done without knowing the attacker's genuine identity. Although the most common goal of this kind of assault is to trick the victim into giving sensitive information—such as login passwords for specific accounts—it may also include other activities, including sending money to the attacker[4][5]. The information provided is voluntarily given by the target of a phishing email assault because they think the message came from a reliable source. Thus, phishing attempts result in the acquisition of passwords or other sensitive data that would have been very challenging to gain in any other way. Filtering email components, senders, URLs, IP addresses, and other data was the foundation of several early systems against phishing attempts. Spam is a category that includes phishing emails. Emails inviting users to click on an embedded link are sent to them by what seem to be reputable companies or banks. The hyperlink should reroute the user to a fraudulent website that solicits private data, such as credit card numbers, usernames, and passwords[6].

The effect of phishing emails on user security has drawn much attention to detecting these emails. Consequently, several methods have been devised to identify phishing emails, ranging from communication-focused approaches like authentication protocols, white-listing, and blacklisting, to content-based filtering techniques [7]. Though they are not widely utilised, blacklisting and whitelisting strategies have not shown themselves to be adequately effective when applied to various domains. Meanwhile, content-based phishing filters are extensively used and have been shown to be quite

effective. Because of this, studies have concentrated on mechanisms that rely on email content, such as creating data mining and ML algorithms that exploit email headers and bodies [8].

This research focuses on Detecting phishing emails, which uses ML methods. ML is a subfield of AI that focusses on teaching computers new skills without having to teach them specifically. Classification in our model is handled using supervised ML algorithms. Supervised learning systems predict unknown data by comparing it to known occurrences. ML algorithms, which are a subset, repeatedly learn from data [9].

Phishing attacks remain a prevalent and dangerous threat to cybersecurity, often leading to financial loss, data breaches, and identity theft. This research aims to enhance cybersecurity by developing and implementing advanced ML techniques for the accurate identification and classification of phishing emails, thus reducing the risk of phishing attacks in digital communication systems. This research aims to address the growing sophistication of phishing tactics by leveraging advanced machine-learning techniques. By developing accurate detection models, such as the hybrid BERT-LSTM, this study aims to significantly reduce the risk of phishing attacks, improve email security, and safeguard sensitive information. The significance lies in providing a scalable, efficient solution that strengthens cybersecurity frameworks and helps organisations and individuals protect themselves from phishing threats.

A. Contribution and aim of paper

Particularly in phishing email detection and categorisation, this work significantly advances cybersecurity. The key contributions are:

- The phishing email dataset for robust model training and evaluation.
- Introduced effective preprocessing techniques, enhancing the quality and relevance of the phishing email dataset for machine learning models.
- Conducted a comparative analysis of SVM, MNB, DT, and BERT-LSTM, showcasing BERT-LSTM's superior performance.
- Employed comprehensive evaluation metrics (F1-score, recall, accuracy, precision) to assess model efficiency for phishing detection.
- Developed a hybrid BERT-LSTM model achieving 99.61% accuracy for phishing email detection, improving cybersecurity defences.
- Applied advanced machine learning to significantly improve phishing email detection, contributing to stronger cybersecurity systems.

B. Structure of paper

This is the outline for the remaining sections of the paper. Section II provides background research on the topic of the classification of phishing emails. Section III lays forth the procedures. Section IV contrasts the results, evaluation, and conversation. The findings and future directions of the study are detailed in Section V.

II. LITERATURE REVIEW

The previous study on Phishing email classification and identification using machine and deep learning approaches is provided in this section.

In, Abadla et al., (2023) To prevent phishing emails from reaching inboxes, an IDS model based on machine learning is suggested. Evaluation scores were created and analysed for several ML classifiers that were put into action, including RF, SVM, Adaboost, LR, and KNN. Two classifiers, Random Forest (98.6% accuracy) and Adaboost (98.1% accuracy) used recursive feature removal and multi-feature analysis to win the competition[10].

In, Almejrab et al. (2023) examined six classifiers to identify the top machine-learning classifiers for phishing attack detection. This work uses CatBoost Classifier (CB), LGBM, XGB, GBA, AdaBoost Classifier and Random Forest Classifier (RF) algorithms. The algorithms have achieved the XGB Classifier's accuracy result of 99.05% and the highest f1-core result of 99.0331%. The outcomes show that XGB Classifier is a reliable classifier for phishing attack detection[11].

In, (Giri et al., (2022) A comparison was made between two DL models' capabilities in identifying phishing emails. The first model makes use of a CNN that incorporates Global Vector (GloVe) word embedding, while the second model employs a BERT model that has been fine-tuned. The suggested approach identifies the phishing email by examining the email's content. Several popular datasets are combined and used to evaluate the model's efficacy. These are lingSpam, Enron Spam Subset, full Spam Assassin, Jose Nazario's phishing dataset, and the Enron email dataset. In this scenario, the GloVe word embedding outperforms the BERT model (98%) in detecting phishing emails[12].

In, Ripa, Islam and Arifuzzaman, (2021) implemented a machine learning-based spear phishing bot on Twitter. With an emphasis on time to train the dataset, we achieved greater accuracy in detecting phishing URLs using several classifiers. Our results show that XGBoost classifier is faster and more accurate (94.44%). They obtained a 95.15% accuracy rate using a NBC to identify phishing emails. After experimenting with several classifiers in our website identification methods, we settled on the RFC, which achieved a remarkable 96.80% accuracy rate[13].

This, Li, Zhang and Wu, (2020) article illustrates how phishing emails use the persuasion principle. Then it builds on previous work to provide a mechanism for detecting these emails that is based on this principle. Whether the matching feature word shows up in the message is a key component of the concept of persuasion. After applying an information gain method to each feature, a final list of 25 features is generated for detection. After extensive testing, the accuracy percentage was finally confirmed to be 99.6 per cent [14].

Table 1 provides the comparative analysis for background study on phishing email detection using machine learning.

Table. 1 Analysis of phishing email detection using ML

Reference	**Methodologies**	**Results**	**Limitations/Future Work**
[10]	Machine learning-based IDS model Classifiers: Random Forest (RF), SVM, Adaboost, Logistic Regression, KNN, Recursive feature elimination and multi-feature analysis	Random Forest achieved highest accuracy of 98.6%,andAdaboost achieved 98.1% accuracy.	Limited focus on classifiers, Explore deep learning models and computational efficiency improvements.
[11]	Classifiers: CatBoost (CB), LightGBM, XGBoost (XGB), Gradient Boosting (GBA), Adaboost, and Random Forest	XGBoost achieved highest accuracy of 99.05% and F1-score of 99.0331%	Limited to machine learning, no feature engineering
[12]	CNN with GloVe word embedding BERT model with fine-tuning	GloVe embedding achieved 98% accuracyand BERT achieved 96% accuracy	Dataset size and variety limitations, limited BERT optimisation Fine-tune BERT, experiment with more diverse datasets.
[13]	Twitter spear phishing bot-URL detection with XGBoost, Naive Bayes for email, and Random Forest for websites	XGBoost achieved 94.44% accuracy (URL detection) and Random Forest achieved 96.80% accuracy (website detection)	Focus on phishing URL detection only, small dataset size.Explore hybrid models, test on larger datasets.
[14]	Phishing email detection using thepersuasion principle Information gain algorithm for feature selection (25 features)	Achieved 99.6% accuracy	Relies on specific feature selection methods Investigate other feature selection techniques, apply to real-time detection.

A. Research Gap

Table 1 provides the summary of related work based on a comparative analysis of performance and methodology. These papers provide helpful information about phishing detection with ML and DL, specifically how models like CNN, RF, and XGBoost may achieve good accuracy. While feature selection approaches can improve performance, the absence of precise datasets in many types of research makes them not very generalisable. Although there is promise in deep learning, it is not yet able to consistently beat more conventional models. There is a need for more comprehensive solutions as most research only addresses certain phishing vectors. Combining deep learning with ensemble techniques and utilising more varied datasets should be explored in future studies.

III. METHODOLOGY

The goal here is to improve cybersecurity by using MLmethods to identify phishing emails. The methodology for identifying and classifying phishing emails consists of several essential stages. Initially, a phishing email dataset is compiled, consisting of 525,754 instances, including 8,351 phishing and 517,402 legitimate emails. Data preprocessing is the subsequent step, which involves the elimination of stop words, HTML tags, and irrelevant numbers to prepare the data. The cleansed text is subsequently transformed into numerical representations through tokenisation during feature extraction. The data is divided into training and test sets, with a typical spread of 70% for training and 30% for testing. Numerous machine learning models, such as SVM, MNB, a BERT-LSTM model, and DT, are implemented to classify the emails. To evaluate the model efficiency with performance matrix including F1-score, recall, accuracy, and precision. This comprehensive approach aims to enhance cybersecurity by accurately identifying and classifying phishing emails. The overall process of classifying phishing emails is displayed in Figure 1 data flow diagram for phishing attack detection.

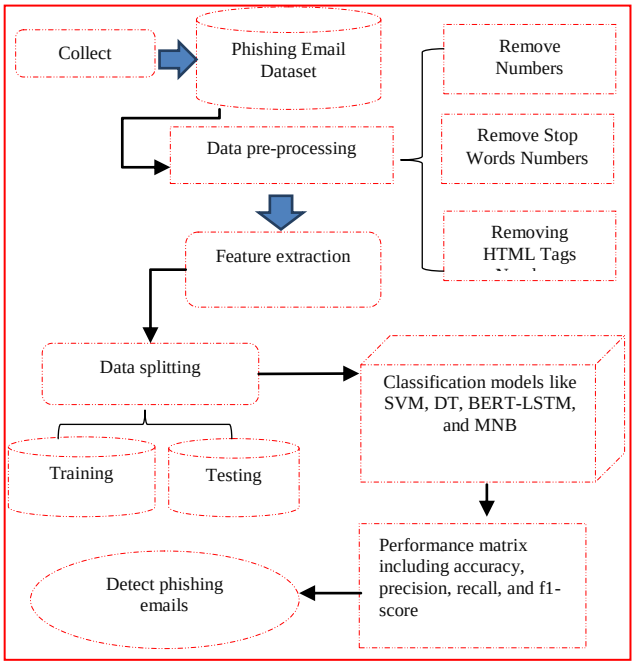

Fig. 1 Methodology flowchart for classifying phishing emails

The overall process of data flow diagram with brief explanation is provided below:

A. Phishing Email Dataset (Collection)

To study and analyse phishing attempts, researchers compile email messages that have been hand-picked or generated for this purpose. This dataset was the first one we ever used; it has 5,25,754 occurrences total, including 5,17,402 legitimate emails and 8351 phishing emails.

B. Data Preprocessing

An ongoing process, data preparation involves transforming raw data into a more comprehensible and usable format. Incomplete, inconsistent, behavior-free, and error-ridden, raw datasets are the norm. It is provided below the data pre-processing steps:

- **Remove Numbers:** It is possible to filter out any numbers or characters that aren't essential to the task of identifying phishing emails.
- **Removing HTML Tags:** Phishing emails may include information that can be found in HTML tags, such as "form" elements that are used to generate a phoney login page. It is possible to keep these pertinent tags and eliminate the others.
- **Remove Stop Words:** Either a stop word filter or a list of stop words may be used to eliminate stop words from the dataset.

C. Feature Extraction

An essential part of text categorisation is feature extraction, which is turning text input into a numerical form that DL models can understand. At its core, this stage is about mining the text for useful information that may inform the classifier's training and evaluation. By tokenising each word in the text, one of the most used methods for feature extraction is achieved.

D. Data Splitting

Partitioning the data into two sets, one for training and one for testing, was the subsequent stage. By using the training dataset, the DL model was trained, and its efficacy was assessed using the test dataset. As an example, imagine a data split of 70/30: 70% for training and 30% for testing.

E. Models

The many MLmodels that were used in the study are detailed in the sections that follow.

1) Support Vector Machine (SVM)

The machine learning technique known as SVM is versatile and may be used for both regression and classification problems. The optimal hyperplane for class separation or regression analysis target value prediction is found using this method [15]. Using a kernel function, SVM expands the dimensionality of the input data and then

optimises a cost function to maximise margin among classes or minimise error for regression.

2) Multi-nomial NB

The Multinomial Naive Bayesian Classifier incorporates multinomially distributed data in addition to being based on the Bayes theorem [16]. For each class y, vectors provide the multinominal distribution, where θyi denotes the likelihood of the feature's appearance in a classy sample and n is the number of features used.

3) Hybrid BERT-LSTM Model

A hybrid BERT-LSTM model leverages the strengths of BERT for contextual language understanding and LSTM for capturing long-term dependencies in sequential data. BERT, with its transformer-based architecture, generates rich, bidirectional contextual embeddings of input text, which are then passed into an LSTM layer [17][18][19]. The LSTM layer processes these embeddings sequentially, retaining important temporal information from the input, which is useful for tasks like sequence classification or prediction. This combination allows the model to utilise BERT's robust language representations alongside LSTM's ability to model time dependencies, resulting in improved performance on tasks [20][21]. Figure 2 shows the BERT-LSTM structure.

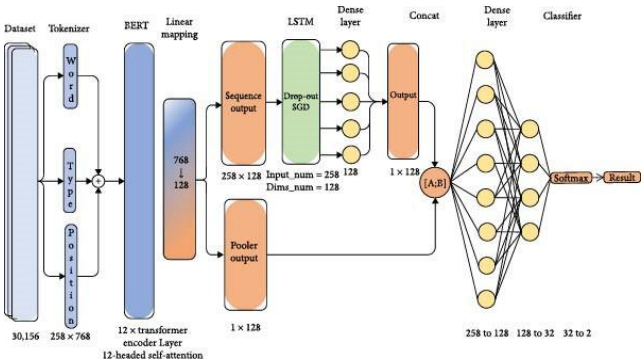

Fig. 2 The structure of BERT-LSTM

4) Decision Tree (DT)

Classification and regression are two applications of decision trees, which are ML algorithms. It builds a DT model iteratively by splitting data into subsets according to a selected characteristic and continuing until a stopping requirement is satisfied. A choice based on a feature is represented by each node, and every branch represents a potential conclusion. Both numerical and categorical data are easily handled by decision trees.

5) Evaluation metrics

In order to assess how well phishing email detection worked, a collection of assessment measures, sometimes called performance metrics, were used. To assess the accuracy of a model, one may utilise a confusion matrix, a Table that compares the model's projected values to the actual values. Accuracy, precision, recall, and F1-score were the five assessment metrics used to assess the final models. Initially, models are evaluated using confusion matrices based on the following: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

- The existence of any unwanted occurrence is represented by the number of records that were appropriately categorised by TP.
- The existence of any undesired occurrence is shown by the number of records that were wrongly categorised by FP.
- The number of records that were successfully categorised is shown as typical in TN.
- The number of records that were misclassified is shown as usual in FN.

Accuracy: The ratio of the number of accurate classifications of normal and undesirable occurrences to the total number of events in the dataset pertaining to oil wells is the accuracy. The expression is (1):

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (1)$$

Precision: The term "precision" refers to the ratio of the number of accurately identified undesirable well events to the total number of unwanted well events. The notation for it is (2):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall: Recall is the ratio of the number of undesirable oil well events in the dataset to the proportion of properly diagnosed unwanted well events. It's shown (3):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1-score: The F1-score, which is formally written as (4), is the weighted average of the recall and precision.

$$F1 = \frac{2*(\text{precision}*\text{recall})}{\text{precision}+\text{recall}} \quad (4)$$

These matrices are used for comparative analysis of machine learning models on phishing email datasets.

IV. RESULTS AND DISCUSSION

This section and its subsections provide the following explanation and analysis of the results. Firstly, provide the result analysis of machine learning models which implement on the phishing email dataset. Also, provides a comparative analysis of various models across the performance matrix.

A. Result analysis

The experiment results analysis of BERT-LSTM models is provided in this section. The results are in the tabular, Figures, and bar graphs forms. The following Table 2 shows the BERT-LSTM model achieving 99.61% accuracy.

Table. 2 Performance of BERT-LSTM model Performance for phishing detection

Measures	**BERT-LSTM**
Accuracy	99.61
Precision	99.87
Recall	99.23
F1-score	99.55

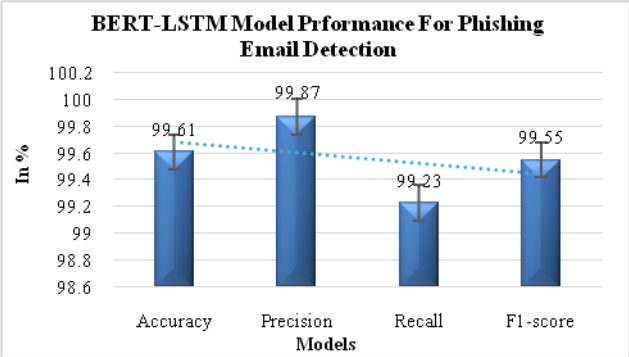

Fig. 3BERT-LSTM model Performance

The above Table 2 and Figure 3 show the BERT-LSTM model demonstrates excellent performance, achieving 99.61% accuracy, 99.87% precision, 99.23% recall, and a 99.55% F1-score, indicating its high accuracy, minimal false positives, and a strong balance between precision and recall.

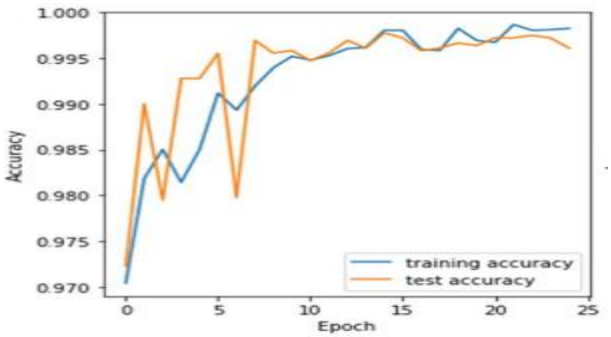

Fig. 4 Training and Testing Accuracy for BERT-LSTM

Figure 4 displays the accuracy of a BERT-LSTM model throughout 25 epochs, both in training and testing. Shown on the y-axis are accuracy values ranging from 0.970 to 1.000, while the x-axis shows the epochs. The training accuracy starts high, briefly drops, and then surpasses the test accuracy, with both lines fluctuating slightly. The test accuracy follows a similar trend, reflecting the model's performance on both datasets.

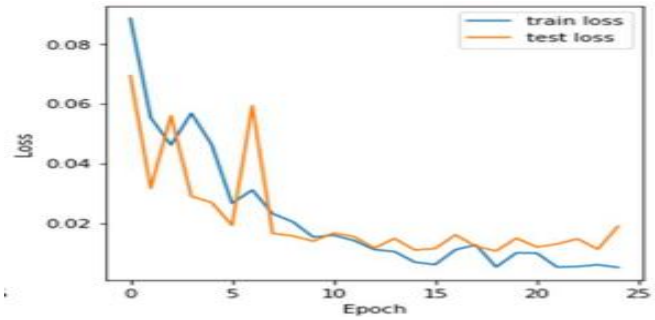

Fig. 5 Training and Testing loss for BERT-LSTM

The blue line stands for training loss and the orange line for testing loss in Figure 5, which shows the BERT-LSTM model's performance throughout 25 epochs. Both lines show a general downward trend, indicating improved model performance as training progresses. However, the testing loss is more variable compared to smoother decline of training loss.

B. Comparative analysis

A comparison between BERT-LSTM and other machine learning models is provided in this section. Table 3 shows BERT-LSTM model outperforms compare to other models.

Table. 3 Comparative analysis of models' performance on Phishing Email Dataset

Models	**Accuracy**	**Precision**	**Recall**	**F1-score**
SVM[22]	81.61	89.74	71.26	79.44
MNB[23]	96.22	96.36	97.36	96.96
DT[24]	96.27	94	96	95
BERT-LSTM	99.61	99.87	99.23	99.55

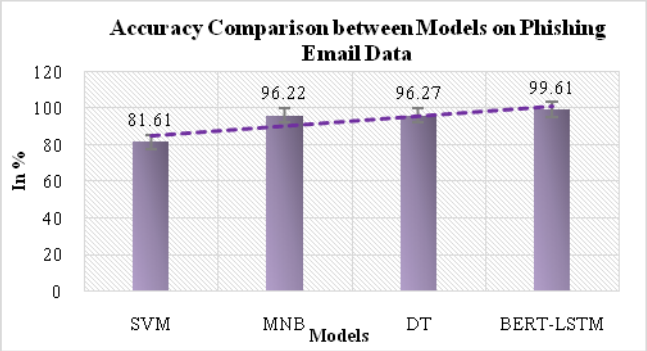

Fig. 6 Bar graph for accuracy comparison of models

Figure 6 displays an accuracy comparison of models. A BERT-LSTM model stands out with a remarkable 99.61%, demonstrating the highest overall correctness in identifying phishing emails. MNB follows closely with 96.22%, and the DT is slightly ahead of MNB at 96.27%. The SVM has a significantly lower accuracy of 81.61%, making it the least accurate among the models. Overall, BERT-LSTM provides the most accurate performance for phishing email detection.

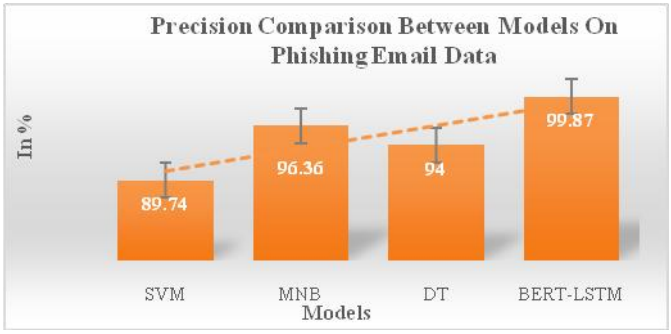

Fig. 7 Bar graph for precision comparison of models

Figure 7 shows a bar graph for the precision comparison of models is shown in Figure 7. The BERT-LSTM model leads in precision with 99.87%, followed by MNB at 96.36%, DT at 94%, and SVM at 89.74%, with BERT-LSTM being the most accurate in minimising false positives.

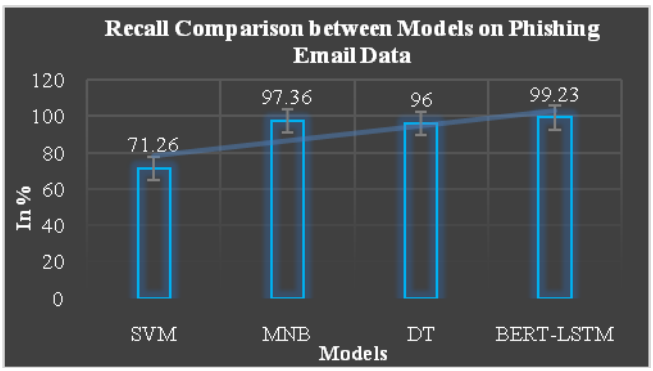

Fig. 8 Bar graph for recall comparison of models

Figure 8 shows a recall comparison of models, where a BERT-LSTM model achieves the highest score at 99.23%, effectively identifying most phishing emails. MNB follows at 97.36%, while DT has a recall of 96%. SVM has the lowest recall at 71.26%, making it less effective in detecting phishing emails.

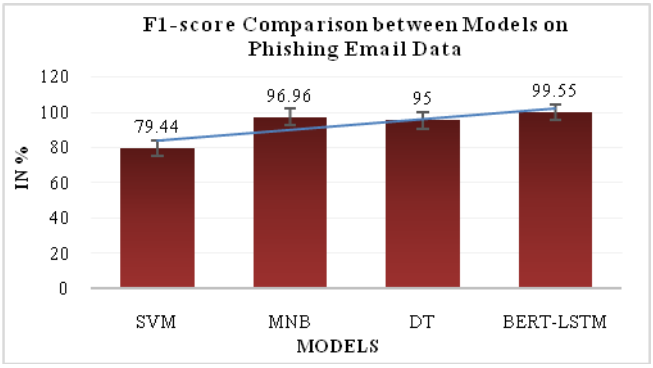

Fig. 9 Bar graph for f1-score comparison of models

Above Figure 9 shows the F1-score; the BERT-LSTM model leads with 99.55, showing an excellent balance between precision and recall. MNB follows at 96.96, with DT close behind at 95. SVM lags with a significantly lower F1-score of 79.44, making it the least effective in balancing precision and recall among the models.

V. CONCLUSION AND FUTURE STUDY

A phishing email is one that seems to be from a trusted source, but is really an attempt by an imposter to trick the recipient into divulging sensitive information. Phishing emails are designed to deceive email users into accessing a malicious website that seems legitimate. Another tactic is to get users to download malicious attachments without realising it. This article offers a concise overview of phishing emails and phishing assaults to help readers get a comprehensive grasp of these types of attacks. In order to identify phishing email attacks, this research compares and contrasts many popular supervised ML methods, including DT, MNB, BERT-LSTM, and SVM. This study reveals that the BERT-LSTM model significantly outperforms other ML models for phishing email detection, demonstrating superior performance with the highest accuracy99.61%, precision99.87%, recall99.23%, and F1-score99.55%. The hybrid BERT-LSTM model excels in understanding contextual information and capturing long-term dependencies, making it the most effective model. The study identifies several limitations and future directions for enhancing phishing email detection. In order to improve the model's performance, future research should concentrate on increasing the dataset's coverage of other types of phishing attacks.

REFERENCES

[1] G. Mohamed, J. Visumathi, M. Mahdal, J. Anand, and M. Elangovan, "An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approach," *Processes*, 2022, doi: 10.3390/pr10071356.

[2] M. Somesha and A. R. Pais, "Classification of Phishing Email Using Word Embedding and Machine Learning Techniques," *J. Cyber Secur. Mobil.*, 2022, doi: 10.13052/jcsm2245-1439.1131.

[3] N. Richardson, V. K. Yarlagadda, S. K. R. Anumandla, and S. C. R. Vennapusa, "Harnessing Kali Linux for Advanced Penetration Testing and Cybersecurity Threat Mitigation," *J. Comput. Digit. Technol.*, vol. 2, no. 1, pp. 22–35, 2024.

[4] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommunication Systems*. 2021. doi: 10.1007/s11235-020-00733-2.

[5] R. P. Vamsi Krishna Yarlagadda, "Secure Programming with SAS: Mitigating Risks and Protecting Data Integrity," *Eng. Int.*, vol. 6, no. 2, pp. 211–222, 2018.

[6] C. S. Eze and L. Shamir, "Analysis and Prevention of AI-Based Phishing Email Attacks," *Electron.*, vol. 13, no. 10, 2024, doi: 10.3390/electronics13101839.

[7] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8] 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.

[9] S. Abraham, "A R eview of Phishing Email Detection based on Different Machine Learning Methods," *Int. J. Eng. Res. Technol.*, vol. 10, no. 04, pp. 240–243, 2022.

[10] R. Abadla, A. Alseiari, A. Alheili, M. S. Daoud, and H. M. Al-Mimi, "Intelligent Phishing Email Detection with Multi-Feature Analysis (IPED-MFA)," in *2023 International Conference on Intelligent Computing, Communication, Networking and Services, ICCNS 2023*, 2023. doi: 10.1109/ICCNS58795.2023.10193714.

[11] R. M. Almejrab, O. M. Sallabi, F. F. Bushaala, A. B. S. Altajori, and A. A. Mohamed, "A Classification Model For Phishing Detection System Based On Machine Learning Algorithms," in *2023 IEEE 11th International Conference on Systems and Control, ICSC 2023*, 2023. doi: 10.1109/ICSC58660.2023.10449834.

[12] S. Giri, S. Banerjee, K. Bag, and D. Maiti, "Comparative Study of Content-Based Phishing Email Detection Using Global Vector (GloVe) and Bidirectional Encoder Representation from Transformer (BERT) Word Embedding Models," in *2022 1st International Conference on Electrical, Electronics, Information and Communication Technologies, ICEEICT 2022*, 2022. doi: 10.1109/ICEEICT53079.2022.9768612.

[13] S. P. Ripa, F. Islam, and M. 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 2021*, 2021. doi: 10.1109/ACMI53878.2021.9528204.

[14] X. Li, D. Zhang, and B. Wu, "Detection method of phishing email based on persuasion principle," in *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020*, 2020. doi: 10.1109/ITNEC48623.2020.9084766.

[15] S. Mathur., "Supervised Machine Learning-Based Classification and Prediction of Breast Cancer," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12(3), 2024.

[16] K. Chubarian and G. Turán, "Interpretability of Bayesian Network Classifiers: OBDD Approximation and Polynomial Threshold Functions," in *ISAIM 2020 - International Symposium on Artificial Intelligence and Mathematics*, 2020.

[17] K. Ullah *et al.*, "Short-Term Load Forecasting: A Comprehensive Review and Simulation Study with CNN-LSTM Hybrids Approach," *IEEE Access*, vol. 12, no. July, pp. 111858–111881, 2024, doi: 10.1109/ACCESS.2024.3440631.

[18] R. Tandon, "Social Media Texts Classification and Analysis Based on Large Language Models (LLMs)," *J. Emerg. Technol. Innov. Res.*, vol. 11, no. 9, pp. a751–a757, 2024.

[19] R. Tandon, "An Analysis Of COVID-19 Tweets Sentiments Based On Large Language Models (Llms)," *Int. J. Res. Anal. Rev.*, vol. 11, no. 3, pp. 319–328, 2024.

[20] H. Sinha, "Predicting Bitcoin Prices Using Machine Learning Techniques With Historical Data," *Int. J. Creat. Res. Thoughts*, vol. 12, no. 8, 2024, doi: 10.3390/e25050777.

[21] U. Ozdil, B. Arslan, D. E. Tacar, G. Polat, and Ş. Ozan, "Ad Text Classification with Bidirectional Encoder Representations," in *Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021*, 2021. doi: 10.1109/UBMK52708.2021.9558966.

[22] A. Mughaid, S. AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. A. Elsoud, "An intelligent cyber security

phishing detection system using deep learning techniques," *Cluster Comput.*, 2022, doi: 10.1007/s10586-022-03604-4.

[23] A. Sa, P. R. Nishantb, S. Baithac, and K Dinesh Kumard, "An Ensemble Classification Model for Phishing Mail Detection," *5th Int. Conf. Innov. Data Commun. Technol. Appl. (ICIDCA 2024)*, 2024, doi: 10.1016/j.procs.2024.03.286.

[24] B. L. V. S. Ram and A. M. Sowjanya, "PHISHING MAIL DETECTION USING MACHINE LEARNING," *Int. J. Creat. Res. Thoughts*, vol. 11, no. 9, pp. 429–440, 2023.