iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73836

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

A Survey on Phishing Email Detection Techniques: Using LSTM and Deep Learning

Annie Jaison J S[1], Halima Sadiya[2], Himashree S[3], M Jomi Maria Sijo[4], Dr. Anitha T G[5]
Department of Computer Science & Engineering, Sapthagiri College of Engineering, Bengaluru, Karnataka, 560057

Abstract: Phishing attacks, often delivered through deceptive emails, remain one of the most dangerous cyber threats, aiming to steal sensitive information such as passwords and financial data. Traditional detection methods like blacklists and rule-based filters struggle to keep up with evolving tactics. This paper surveys recent deep learning approaches to phishing email detection, focusing on models such as Long Short-Term Memory (LSTM), Convolutional Neural Network(CNN), Bidirectional Long Short-Term Memory (BiLSTM), Transformer encoders, and hybrid architectures. These methods analyze email components like subject lines, content, metadata, and user behavior. The study also reviews commonly used datasets, feature extraction techniques, and evaluation metrics, providing insights into current trends, strengths, and challenges in developing effective phishing detection systems.

I. INTRODUCTION

Phishing is a deceptive cyber-attack method used by malicious actors to obtain sensitive information. These attacks typically appear in the form of legitimate-looking emails, which redirect users to fraudulent websites or prompt them to download malicious content. As phishing techniques continue to evolve, they present a growing threat to individuals and organizations across various domains. With the increasing reliance on digital communication, phishing has emerged as one of the most prevalent and damaging cybersecurity challenges.

Traditional detection methods, such as blacklists or signature-based systems, are no longer sufficient. These approaches struggle to keep pace with the rapidly changing tactics employed by attackers. As a result, there is a pressing need for more intelligent and adaptable detection systems. The use of advanced machine learning and Deep Learning(DL) models has shown promising potential in this area. In particular, models capable of understanding language patterns can significantly enhance the accuracy of phishing detection and help mitigate cyber risks.

This paper focuses on the application of LSTM networks for phishing email detection. LSTMs are a type of deep learning model designed to analyze sequential data, making them well-suited for processing textual content. By learning from the structure, context, and linguistic features of emails, LSTMs can effectively distinguish between legitimate and phishing messages.The study also presents a comparative analysis between LSTM-based models and traditional machine learning approaches such as Naïve Bayes. The results highlight the advantages of deep learning in capturing complex patterns and improving detection performance.

In addition, the paper discusses real-world examples, practical challenges, and important points to consider when building phishing detection systems. It aims to help researchers and developers better understand the current methods and create smarter, easier-to-use security solutions in the future.

II. BACKGROUND

A. Problem Description

Cybersecurity is one of the most critical and complex fields in the digital era. It includes several areas such as network protection, data privacy, malware detection, and email-based threat mitigation. As cyber threats continue to evolve rapidly, traditional rule-based learning approaches have become insufficient for building robust and timely defense mechanisms . Although existing techniques help to detect the system vulnerabilities, they alone do not guarantee effective response. The development of cognitive skills such as analyzing, interpreting, and acting upon security data is equally important and essential for the proactive defense [1].

Among the wide range of defense strategies, artificial intelligence (AI), especially deep learning (DL), has emerged as a key technology in cybersecurity. These approaches can handle large-scale data and extract patterns with high speed and accuracy [2]. DL systems are increasingly applied to detect threats like phishing, malware, and intrusions by identifying abnormal behaviours and adapting to new patterns [3]. For instance, DL models can spot phishing emails by recognizing subtle inconsistencies in structure and language that are hard to catch using traditional techniques [4].

Recent studies also indicate a shift in attacker strategies, with adversaries now leveraging generative AI to create convincing phishing emails. This undermines the effectiveness of signature-based and rule-driven detection methods [5]. Attackers continuously refine their methods—altering URLs, payloads, and message formats—to bypass conventional defenses [6]. To address this, researchers are turning to adversarial-aware models that can anticipate and learn from evolving threats in real time.

B. Adversary Model

In this context, the concept of an adversary model has become increasingly important. It represents the attacker's objectives, tactics, and behaviors, which can range from phishing and malware injection to social engineering and prompt manipulation attacks [7].

DL models—like CNNs, BiLSTMs, and Transformers—can model both sequential and spatial patterns effectively, enabling detection of such attacks with improved precision [8].

Furthermore, emerging work explores hybrid deep learning architectures, federated learning, and behavior-driven analysis, offering better adaptability and user-centric privacy features [9].

As phishing threats grow more complex, these approaches can enhance detection accuracy without compromising user privacy or performance. Studies have also explored real-world deployment considerations, such as reducing model size for mobile applications and increasing robustness against adversarial samples [10].

III. LITERATURE SURVEY

A. LSTM-Based Email Content Classification

Several approaches have adopted LSTM networks to classify emails as phishing or legitimate. These models process textual content including the subject line, body, and embedded URLs. They are particularly effective in modeling sequential dependencies in textual data, which are common in phishing email patterns. The use of LSTMs has reported high accuracy (up to 96–97%) on public datasets like Enron, Nazario, and PhishTanks. However their performance may drop when dealing with short, ambiguous, or heavily obfuscated messages.

B. Hybrid CNN-LSTM Models for Real-Time Detection

Hybrid models that combine CNNs with LSTMs have been explored for real-time phishing detection in varied environments. These models extract both local features (e.g., keyword patterns) and sequential structures in the text. It offers high detection accuracy with deployment feasibility on lightweight or mobile platforms. In contrast, the dual architecture increases model complexity, which can delay the inference on limited hardware.

C. Multimodal Phishing Detection

Multimodal frameworks that include both Visual Textual datas combine (Natural Language Processing) NLP techniques with visual analysis, such as analyzing screenshots or rendered email content. This dual approach helps to detect phishing by analyzing both the words used and the way the email looks. It identifies phishing indicators across both content and visual design, increasing detection robustness. On the other hand, it requires higher computational resources and additional preprocessing steps such as image rendering.

D. Lightweight and Privacy-Preserving Detection on Edge Devices

To address privacy and resource constraints, lightweight LSTM models are being adapted for mobile and edge environments. This ensures data privacy while enabling adaptive learning from distributed sources. It allows secure, personalized detection without compromising user data privacy. That said, training efficiency may be limited by device hardware and the availability of local data.

E. Comparison of LSTM, BiLSTM, and Transformer Architectures

Comparative studies have evaluated the effectiveness of LSTM, BiLSTM, and Transformer-based models in phishing detection. LSTM and BiLSTM perform well in low-resource or noisy environments with shorter inputs. Transformers, on the other hand, offer superior performance on large-scale datasets with complex or lengthy input sequences. It enables architecture selection based on task requirements, available resources, and input size. Nevertheless, transformers require significantly more computation and data during training.

F. Email Header and Metadata Analysis

LSTM models are often enhanced by including email metadata such as headers, domain reputation, and sender history. These features provide structured context that complements unstructured body text. The fusion of both types of data increases the model's ability to detect deceptive behaviour.

It improves detection accuracy by incorporating behavioural and structural features. Even so, it requires access to and processing of additional metadata streams beyond the email content.

G. User Behavior-Based Detection Models

Behavioural modelling uses sequential data like browsing history, and interaction timing to detect phishing.

LSTM networks are employed to learn user behaviour over time and detect deviations. It is effective in identifying phishing through anomalous behavior patterns independent of the email text. Nonetheless, its implementation depends on continuous monitoring and collection of behavioral data which may raise privacy or ethical concerns.

H. Attention Mechanism in LSTM Architectures

Attention layers are added to LSTM architectures to allow the model to focus on the most relevant parts of the email text. This selective attention improves the network's ability to identify phishing indicators. It boosts performance and transparency by directing focus to phishing-relevant tokens, still it adds computational complexity and requires careful fine-tuning for optimal results.

I. Comparison with Traditional Machine Learning

LSTM-based models have consistently outperformed classical Machine Learning(ML) algorithms such as Naive Bayes, and Random Forest in phishing detection. They are particularly better in handling variable-length input. They also adapt more effectively to changes in phishing tactics over a period of time. It demonstrates higher accuracy, recall, and adaptability compared to traditional approaches. But it requires more training data and also results in longer model development cycles.

J. Robustness Against Adversarial Attacks

To counter challenging inputs, some LSTM-based models are trained on modified data to improve resistance against evasion tactics. These models learn to detect subtle changes in phishing content designed to fool detection systems.

It enhances resilience to modified phishing content through adversarial training. Nonetheless, it increases model complexity and requires access to carefully designed adversarial examples during training

These insights help guide the design of stronger and more flexible phishing detection systems.

IV. METHODOLOGY

A. Literature Selection Criteria

Relevant research articles were selected based on the following criteria:

Focus on phishing detection using deep learning

Use of models such as LSTM, CNN, BiLSTM, Transformers, or hybrid approaches

Publication in peer-reviewed journals, conferences, or recognized preprint repositories (e.g., IEEE, Springer, Elsevier, arXiv)

Use of real-world datasets (e.g., Enron, Nazario, PhishTank, Kaggle)

B. Data Sources

The papers were gathered from digital libraries and research databases such as:

IEEE Xplore

SpringerLink

ScienceDirect

Google Scholar

arXiv.org

Keywords used included:

"phishing email detection", "LSTM phishing", "deep learning phishing detection", "email classification CNN LSTM", "transformer-based phishing detection".

C. Categorization of Models

After selection, the models were categorized based on their architectural strategies:
LSTM-only models
CNN-LSTM hybrid models
Attention-based LSTM models
Transformer-based models
Multimodal (text + visual) models
Federated and edge-device models
Behavioral-based models

Each category was studied for:
Architectural design
Datasets used
Feature extraction methods
Evaluation metrics (Accuracy, Precision, Recall, F1-Score)Practical deployment considerations

D. Comparative Analysis

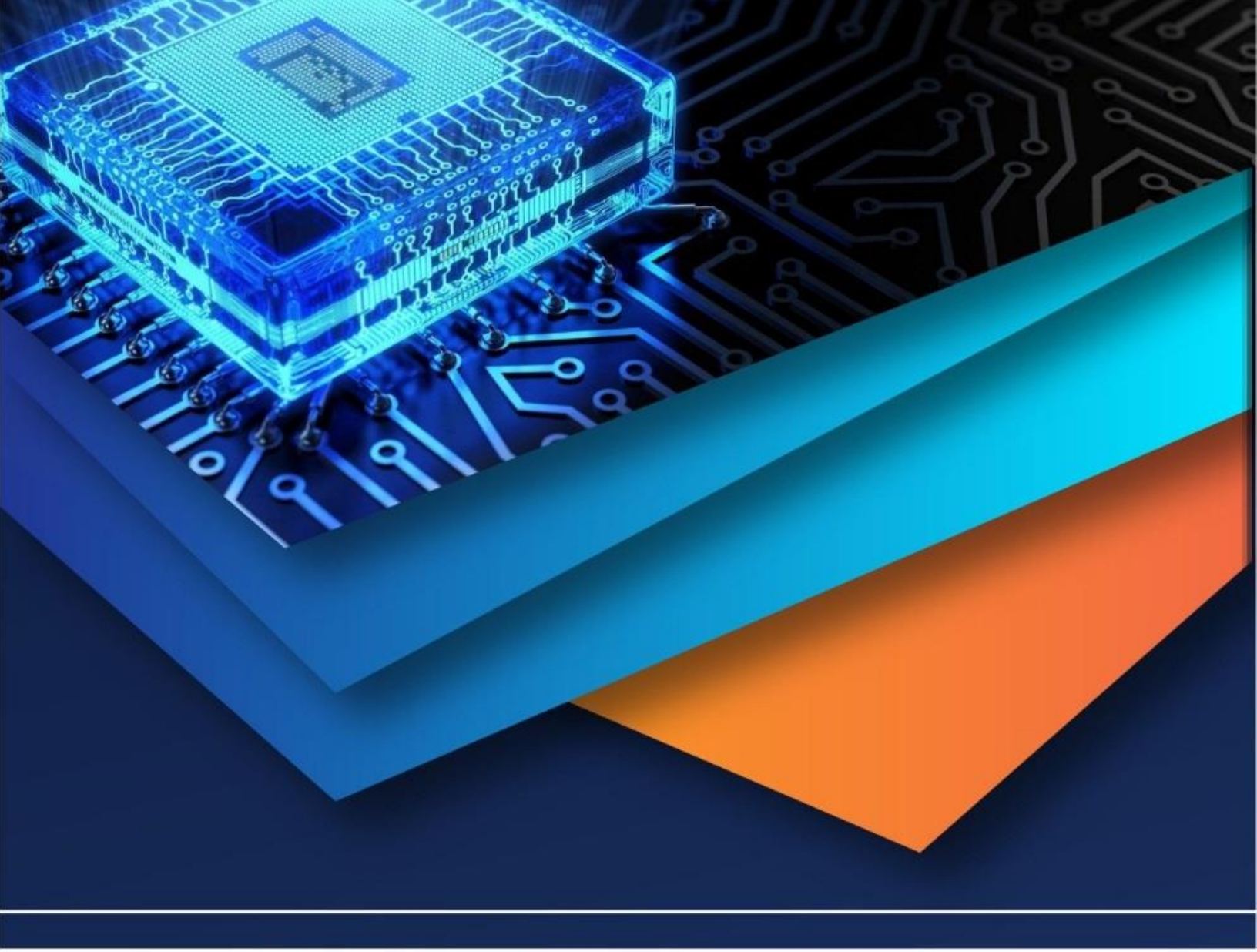
A qualitative analysis was carried out across models to understand:
Common strengths and trends (e.g., LSTM's ability to capture sequence-based features)
General limitations (e.g., computational overhead, lack of robustness to adversarial emails)
Use-case suitability (e.g., mobile, real-time, privacy-preserving detection)

V. CONCLUSION

Phishing continues to be one of the most widespread and dangerous cyber threats, often exploiting human trust to steal sensitive information. This survey explored how Deep Learning, especially models like LSTM, BiLSTM, and Attention mechanisms, is transforming phishing detection by enabling systems to learn patterns, adapt to new tactics, and improve accuracy.
As a future direction, we propose a hybrid phishing detection method that combines TF-IDF-based subject analysis with BiLSTM and Attention for the email body. While this model is yet to be implemented, it draws on strengths observed across existing approaches and presents a strong potential for detecting phishing attempts more effectively by capturing both shallow textual features and deeper contextual patterns. Looking ahead, integrating advanced phishing detection into real-time email systems can greatly improve accuracy and resilience. As phishing strategies continue to evolve in complexity, the need for adaptive, intelligent models becomes increasingly critical. This survey serves as a foundational guide for researchers and practitioners by consolidating existing Deep Learning-based approaches, highlighting their capabilities and limitations, and pointing towards promising directions. Future efforts may focus on scalable deployment, adversarial robustness, and privacy-preserving mechanisms to develop more effective, context-aware, and trustworthy email security solutions.

REFERENCES

[1] S. Baskota, "Phishing URL Detection using Bi-LSTM," arXiv preprint arXiv:2504.21049, 2025.
[2] R. Achary, S. N. Bugath, G. Chakrapani, and M. Venkatesh, "Enhanced Phishing Detection Using LSTM, CNN, and SVM Techniques," in ICTCS 2024, Springer, 2025, pp. 185–204.
[3] H. Malik, P. Awasthi, and R. Sharma, "Deep Learning for Cybersecurity: Threat Detection and Prevention in Complex Networks," Journal of Network and Computer Applications, vol. 229, 103134, 2024.
[4] T. Nguyen, H. Nguyen, and Q. Nguyen, "Email Phishing Detection Using BERT and Transfer Learning," IEEE Access, vol. 10, pp. 105421–105432, 2022.
[5] Y. Zhu and J. Lin, "A Survey on Generative AI for Cyber Threats," Computers & Security, vol. 123, 102942, 2023.
[6] O. Christou, N. Pallis, and G. Pallis, "Phishing URL Detection Through Top-level Domain Analysis," arXiv preprint arXiv:2005.06599, 2022.
[7] M. Sravanth, K. N. Rao, A. Gupta, and R. Raj, "Adversarial Learning for Secure AI Systems," ACM Transactions on Privacy and Security, vol. 26, no. 1, 2023.
[8] K. Zhou, L. Wang, and Z. Zhang, "Advanced Malware and Phishing Detection Using CNN-LSTM Hybrid Models," Computers & Security, vol. 126, 102972, 2023.
[9] P. Singh, R. Dey, and S. Bose, "Federated Learning for Email Threat Detection with Privacy Preservation," IEEE Transactions on Dependable and Secure Computing, 2024.
[10] L. Chan, A. Y. Lee, and M. T. Ho, "Deploying Lightweight Deep Models for Real-Time Phishing Detection," Expert Systems with Applications, vol. 213, 119008, 2024.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY