Detection and prevention of Malicious Urls and phishing attacks using LSTM Model

Panem Aditya[1], Mr.Ch.Raja Jacob[2], Mrs.K.Nirusha[3], Mr.P.Mareswara Rao[4]

[1]PG student ,Department Of Computer Science & Engineering, Mother Teresa Institute Of Science And Technology Autonomous, Sanketika Nagar, Kothuru (V), Sathupally - 507303, Khammam Dist., Telangana,India

[2,3,4]Assistant Professor , Department Of Computer Science & Engineering, Mother Teresa Institute Of Science And Technology Autonomous, Sanketika Nagar, Kothuru (V), Sathupally - 507303, Khammam Dist., Telangana,India

***Abstract*—**The rapid growth of internet services and online communication has increased the number of cyber threats targeting individuals and organizations. Among these threats, malicious URLs and phishing attacks are considered highly dangerous because they deceive users into revealing confidential information such as passwords, banking details, and personal data. Traditional security systems based on blacklists and rule-based detection methods are becoming less effective due to the continuously evolving nature of phishing techniques. Attackers generate new URLs and phishing websites dynamically, making it difficult for conventional machine learning approaches to identify unknown threats accurately. Therefore, there is a growing need for intelligent and adaptive systems capable of detecting malicious URLs in real time with higher accuracy and lower false-positive rates.

This project proposes a deep learning-based approach using the Long Short-Term Memory (LSTM) algorithm for the detection and prevention of malicious URLs and phishing attacks. LSTM is a specialized recurrent neural network capable of learning sequential patterns and long-term dependencies in textual data. Since URLs contain sequential character patterns and hidden structures, LSTM can effectively analyze these patterns to distinguish between legitimate and malicious websites. The proposed system collects URL datasets from multiple sources, preprocesses the data, extracts meaningful features, and trains the LSTM model to classify URLs into safe or malicious categories.

I. INTRODUCTION

The internet has become an essential part of modern life, enabling communication, online banking, e-commerce, education, healthcare, and social networking. While internet technologies provide numerous benefits, they also create opportunities for cybercriminals to launch various types of attacks. One of the most common cyber threats today is phishing attacks performed through malicious URLs. Phishing attacks are fraudulent attempts to trick users into revealing sensitive information by directing them to fake websites that imitate legitimate organizations. These attacks cause severe financial losses, identity theft, data breaches, and privacy violations worldwide.

Malicious URLs are web links designed to distribute malware, steal user credentials, redirect users to harmful websites, or exploit system vulnerabilities. Cyber attackers

Doi:10.48047/ijiee.2026.16.2.48

continuously modify their techniques to bypass traditional security mechanisms. Conventional methods such as blacklist filtering, signature-based detection, and rule-based systems are unable to effectively detect newly generated phishing URLs because these approaches rely heavily on predefined patterns and previously known threats. As attackers rapidly generate new malicious domains and obfuscate URL structures, the need for intelligent and adaptive detection systems has become increasingly important.

II. LITERATURE SURVEY

1. Malicious URL Detection using Machine Learning: A Survey

Authors: Doyen Sahoo, Chenghao Liu, Steven C. H. Hoi

- **Technique Used:** Machine Learning algorithms for malicious URL classification and detection.
- **Pros:**
  - Detects unknown malicious URLs.
  - Improves cybersecurity protection.
  - Supports automated detection systems.
- **Cons:**
  - Requires large datasets.
  - Performance depends on feature selection.

2. A Systematic Literature Review on Phishing Website Detection Techniques

Authors: Asadullah Safi, Satwinder Singh

- **Technique Used:** Blacklist-based, heuristic, machine learning, and deep learning methods.
- **Pros:**
  - Comprehensive comparison of phishing detection techniques.
  - Identifies effective ML-based approaches.
- **Cons:**
  - Traditional blacklist methods fail against new attacks.
  - High computational requirements for deep learning.

3. A Comprehensive Literature Review on Phishing URL Detection Using Deep Learning Techniques

Authors: Kritika

- **Technique Used:** Deep learning models including CNN, RNN, and LSTM for phishing URL detection.
- **Pros:**
  - High detection accuracy.
  - Learns hidden URL patterns automatically.
- **Cons:**
  - Requires extensive training data.
  - Increased training complexity.

4. Phishing Website Detection: A Systematic Review of URL-Based and Deep Learning Approaches

Authors: Pratik Ahirwar, Anurag Shrivastava

- **Technique Used:** URL feature extraction and deep learning-based phishing detection.
- **Pros:**
  - Effective real-time detection.
  - Improved classification accuracy.
- **Cons:**

Doi:10.48047/ijiee.2026.16.2.48

- False positives may occur.
- Requires frequent model updates.

5. Intelligent Detection Designs of HTML URL Phishing Attacks

Authors: Sk. Nishanth Anjum, Mohammad Rahmat Ali, D.V.S.S. Subramanyam

- **Technique Used:** HTML analysis, URL feature extraction, and machine learning techniques.
- **Pros:**
  - Detects sophisticated phishing URLs.
  - Supports hybrid detection methods.
- **Cons:**
  - Complex preprocessing steps.
  - Higher computational overhead.

III. EXISTING SYSTEM

Existing systems for malicious URL detection primarily rely on traditional machine learning algorithms such as Decision Tree, Naive Bayes, Random Forest, Support Vector Machine (SVM), Logistic Regression, and K-Nearest Neighbor (KNN). These systems attempt to classify URLs as legitimate or malicious by extracting handcrafted features from URLs and website content. Common features include URL length, number of special characters, domain age, suspicious keywords, IP address usage, subdomain count, and abnormal redirection patterns. These extracted features are then used to train machine learning models capable of identifying phishing websites and harmful links.

Traditional machine learning-based phishing detection systems operate by first preprocessing URL datasets and manually selecting relevant features that represent malicious behavior. After feature extraction, classification algorithms are trained using labeled datasets containing both safe and malicious URLs. During prediction, the trained models analyze incoming URLs and classify them based on learned feature patterns. Many cybersecurity applications and email filtering systems currently use these ML-based detection techniques because they provide faster processing and moderate detection accuracy.

Disadvantages

Existing machine learning-based systems for malicious URL detection and phishing prevention suffer from several limitations that reduce their effectiveness against modern cyber threats. One of the major disadvantages is the heavy dependency on manual feature extraction. Traditional machine learning algorithms such as Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine require cybersecurity experts to manually identify and engineer relevant URL features before model training. This process is time-consuming, complex, and highly dependent on domain expertise. If important features are missed during extraction, the detection accuracy of the system decreases significantly.

Another major limitation of existing systems is their inability to detect zero-day phishing attacks effectively. Cyber attackers continuously generate new malicious URLs using advanced obfuscation techniques, random domain names, shortened links, and dynamically changing structures. Traditional machine learning models rely on previously learned patterns and handcrafted features, making them less adaptable to newly emerging attack strategies. As a result, unknown phishing websites may bypass security detection systems and compromise user information.

Doi:10.48047/ijiee.2026.16.2.48

IV. PROPOSED SYSTEM

The proposed system introduces a deep learning-based approach using the Long Short-Term Memory (LSTM) algorithm for the detection and prevention of malicious URLs and phishing attacks. The system is designed to overcome the limitations of traditional machine learning techniques by automatically learning hidden sequential patterns from URL data without relying heavily on manual feature engineering. LSTM is a specialized form of recurrent neural network capable of processing sequential information and remembering long-term dependencies, making it highly effective for analyzing textual URL structures.

The proposed system begins with collecting a large dataset containing both legitimate and malicious URLs from trusted cybersecurity repositories and phishing databases. These datasets undergo preprocessing steps such as duplicate removal, normalization, tokenization, sequence conversion, and padding. Since URLs are textual sequences, the preprocessing stage transforms them into numerical representations that can be understood by the LSTM network.

Advantages

The proposed LSTM-based malicious URL detection and phishing prevention system offers several significant advantages over traditional machine learning and rule-based cybersecurity approaches. One of the primary advantages is its ability to automatically learn complex sequential patterns directly from raw URL data. Unlike conventional machine learning algorithms that require manual feature extraction, the LSTM model independently identifies hidden malicious characteristics within URL structures. This reduces dependency on human expertise and improves overall detection efficiency.

Another major advantage of the proposed system is its capability to detect zero-day phishing attacks and previously unseen malicious URLs. Cyber attackers constantly modify phishing strategies by creating new domain structures, random character combinations, and deceptive subdomains to bypass existing security systems. Traditional detection methods struggle to identify these newly generated threats because they rely heavily on predefined rules or historical patterns. In contrast, the LSTM algorithm learns generalized behavioral patterns from training data, enabling it to recognize unknown phishing attacks with higher accuracy.

V. SYSTEM ARCHITECTURE

The overall architecture of the phishing detection system is divided into multiple layers to improve modularity, scalability, and efficiency. These layers include the data acquisition layer, preprocessing layer, feature engineering layer, deep learning layer, prediction layer, prevention layer, and user interaction layer. Each layer communicates with the next layer through secure data pipelines.

VI. RESULTS AND OUTCOMES

Doi:10.48047/ijiee.2026.16.2.48

Screen 1: Home Page

Screen 2: Use login Page

Screen 3: Accuracy Page

VII. CONCLUSION

The project titled "Detection and Prevention of Malicious URLs and Phishing Attacks Using LSTM Model" demonstrates the importance of intelligent cybersecurity systems in protecting users from modern web-based threats. With the rapid growth of internet services, online banking, social media platforms, cloud applications, and e-commerce websites, phishing attacks and malicious URLs have become one of the most dangerous cybersecurity issues. Traditional blacklist-based and rule-based detection methods are often unable to identify newly generated phishing websites because attackers continuously change domain names, URL structures, and attack patterns. To overcome these limitations, the proposed system utilizes a Long Short-Term Memory (LSTM) deep learning model to identify malicious URLs based on sequential URL patterns and hidden textual relationships.

The implemented LSTM model successfully analyzes URL structures, lexical patterns, suspicious keywords, domain behavior, and sequence dependencies to classify URLs as either legitimate or malicious. Unlike conventional machine learning techniques that require manual feature engineering, the LSTM model automatically learns important features from URL sequences. This improves detection accuracy and enables the system to identify complex phishing patterns more effectively. The experimental analysis shows that the proposed model achieves high accuracy, precision, recall, and F1-score while reducing false positive rates. The ability of the LSTM network to retain long-term dependencies helps the system recognize subtle malicious characteristics that are difficult to detect using traditional algorithms.

VIII. BIBLIOGRAPHY

[1] S. Sahoo, B. B. Gupta, and S. Misra, "Detection of phishing websites using machine learning techniques," *Journal of Network and Computer Applications*, vol. 145, pp. 102–118, Oct. 2019.

[2] A. Jain and B. Gupta, "Phishing detection: Analysis of visual similarity based approaches,"

Doi:10.48047/ijiee.2026.16.2.48

Security and Communication Networks, vol. 10, no. 13, pp. 2168–2184, 2017.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[5] R. Verma and A. Das, "What's in a URL: Fast feature extraction and malicious URL detection," in *Proc. 3rd ACM Cyber-Physical System Security Workshop*, Abu Dhabi, UAE, 2017, pp. 55–63.

[6] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009, pp. 1245–1254.

[7] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, Oct. 2014.

[8] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, Fourth Quarter 2013.

[9] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," in *Proc. 11th ACM Symposium on Document Engineering*, Mountain View, CA, USA, 2011, pp. 259–262.

[10] B. B. Gupta, N. A. G. Arachchilage, and K. E. Psannis, "Defending against phishing attacks: Taxonomy of methods, current issues and future directions," *Telecommunication Systems*, vol. 67, no. 2, pp. 247–267, Feb. 2018.

Doi:10.48047/ijiee.2026.16.2.48