Interpretable Deep Learning-Based Phishing Detection Using Convolutional and LSTM Networks

Dewashish Kumar
CSE, Lovely Professional University, Phagwara, Punjab, India

Amanpreet Kaur
CSE, Lovely Professional University, Phagwara, Punjab, India

Mohit Dewangan
CSE, Lovely Professional University, Phagwara, Punjab, India

Mahipal Singh Papola
CSE, Lovely Professional University, Phagwara, Punjab, India

Abstract

Phishing continues to be a major cybersecurity concern, as attackers increasingly rely on deceptive websites and social engineering techniques to obtain sensitive user information. Accurately identifying phishing websites remains difficult because attack patterns evolve rapidly and web-related features are often complex and interdependent. In this work, we develop an explainable deep learning approach for phishing website detection based on a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. The model is trained and evaluated using the UCI Phishing Websites Dataset, which contains a diverse set of URL and HTML-related features. CNN layers are used to capture local structural patterns in the input features, while LSTM units model sequential relationships that are difficult to learn using conventional classifiers. Experimental evaluation shows that the proposed model achieves an accuracy of approximately 95–96%, outperforming traditional machine learning approaches such as Random Forest. To improve transparency, SHapley Additive exPlanations (SHAP) are employed to analyze model decisions and highlight the contribution of individual features. The explainability analysis indicates that factors such as URL length, redirection behavior, and the use of special characters play a key role in phishing classification. By combining strong detection performance with meaningful explanations, the proposed framework offers a practical and trustworthy solution for real-world phishing detection systems.

Keywords

Phishing website detection, explainable artificial intelligence, hybrid CNN–LSTM architecture, convolutional neural networks, long short-term memory networks, SHAP-based interpretability, deep learning methods, UCI phishing websites dataset, cybersecurity analytics, web security.