CITE PUBLICATIONS

Journal of Analytical and Applied Computer Sciences
Volume: 2, Issue: 2, 2026 (July - December), Pages: 32-37
ISSN (Online): 3139-3349 | DOI: https://doi.org/10.67914/jaacs/rsa/2.2/32-37
Research Article

An Intelligent Intrusion Detection Model for Prevention of Social Engineering Attacks on Social Media Network Platforms Using CNN-RNN-LSTM

Sulaiman Ahmad*, Peter Buba Zirra

Department of Computer Science, Faculty of Science, Federal University Kashere, Gombe State, Nigeria.

***Correspondence**
Sulaiman Ahmad
abudalha1023@gmail.com

Received: 21 April 2026 / Accepted: 30 July 2026 / Published: 2 September 2026

The increasing use of social media network platforms has created an attractive environment for social engineering attacks, where attackers exploit human trust, fear, urgency and other psychological cues rather than relying exclusively on technical vulnerabilities. This study presents an Intelligent Intrusion Detection Model (IIDM) for detecting social engineering attacks in textual social media content using a hybrid Convolutional Neural Network–Recurrent Neural Network–Long Short-Term Memory (CNN-RNN-LSTM) architecture. Natural Language Processing (NLP) techniques were applied to prepare and represent social media text, while the hybrid architecture was designed to combine local linguistic feature extraction with sequential/contextual pattern modelling. A cross-platform dataset containing 20,000 benign and malicious social media posts from Facebook and Twitter was cleaned, normalized and tokenized, with Synthetic Minority Over-sampling Technique (SMOTE) used to address class imbalance. The dataset was divided into 80% training and 20% testing sets. The model was trained using categorical cross-entropy and the Adam optimizer, with batch normalization, dropout and early stopping used to improve generalization. Experimental results showed an overall accuracy of 95%, macro precision of 0.95, macro recall of 0.89, macro F1-score of 0.91 and ROC-AUC of 0.89. Class-wise analysis showed strong benign classification, with precision of 0.95, recall of 0.99 and F1-score of 0.97, while the malicious class achieved precision of 0.94, recall of 0.78 and F1-score of 0.85. The results indicate that the hybrid model provides a practical approach for detecting social engineering threats in unstructured social media text and improves upon the social-media baseline reported in the study.

Keywords: *Social engineering, intrusion detection, social media, natural language processing, CNN; RNN-LSTM, deep learning, phishing*

Introduction

Social media network platforms have become major channels for communication, information sharing, marketing and social interaction. Their broad reach and the large volume of user-generated content have also increased opportunities for cybercriminals to exploit users and their trusted relationships. Unlike attacks that depend mainly on software or hardware vulnerabilities, social engineering attacks target human judgement by manipulating trust, fear, curiosity, urgency and perceived authority. On social media, such attacks may appear as phishing messages, impersonation, clickbait, scareware, fake requests or other deceptive interactions.

Traditional security mechanisms such as password authentication, role-based access control and conventional intrusion detection remain important components of information security, but they have limitations when an attacker uses legitimate or compromised credentials. The dissertation identifies this problem as particularly important for social media because a compromised account can be used to deceive the victim's contacts and distribute further malicious content. Existing approaches also face challenges associated with false alarms, heterogeneous user-generated data and limited generalization when models are developed for a single platform. Previous work reviewed in the study includes machine-learning and deep-learning approaches for intrusion detection, phishing detection, fake-account detection and social-media security. A particularly relevant social-engineering study by Adekunle et al. (2024), as reported in the dissertation, used an RNN-LSTM model on Facebook posts and achieved precision of 0.85 and recall of 0.79. The dissertation identifies limitations in single-platform coverage and detection performance as important motivations for a cross-platform hybrid architecture. The present study therefore develops an Intelligent Intrusion Detection Model based on CNN-RNN-LSTM. CNN layers are used to identify local textual patterns, phrases and spatial relationships, while RNN-LSTM layers model longer-range sequential dependencies. NLP is incorporated to transform social-media text into representations suitable for deep-learning classification. The study focuses on textual data from Facebook and Twitter and evaluates the model using accuracy, precision, recall, F1-score and ROC-AUC.

Problem Statement

The increasing popularity of social media platforms has made them attractive targets for cybercriminals seeking to exploit users through social engineering attacks such as phishing, impersonation, scareware, and clickbait. Unlike conventional cyberattacks that primarily exploit technical vulnerabilities, social engineering attacks exploit human trust and psychological behavior, making them difficult for traditional authentication and intrusion detection mechanisms to identify. Existing security approaches may also suffer from high false-positive rates, limited real-time capability, and reduced effectiveness against sophisticated and evolving attacks.

Existing research has also demonstrated limitations in cross-platform generalization. For instance, the Social Engineering Assault Detection (SEAD) framework reported in the study was developed primarily using Facebook data and an RNN-LSTM architecture, thereby limiting its applicability to other social media environments with different communication patterns and user behaviors. The study further identifies room for improvement in detection performance, particularly for complex and nuanced social engineering attacks.

Therefore, there is a need for a more robust and intelligent approach capable of analyzing social-media text and identifying malicious content across different platforms. This study addresses this gap by developing a hybrid CNN-RNN-LSTM model that combines local textual feature extraction with sequential and contextual feature learning to improve the detection of social engineering attacks on Facebook and Twitter.

Aim and Objectives of the Study

- To design a CNN-RNN-LSTM hybrid model for detecting social engineering threats such as phishing, pretexting, clickbait and scareware.
- To evaluate the performance of the model using accuracy, precision, recall and F1-score.
- To compare the performance of the CNN-RNN-LSTM hybrid model with existing works.

Related Work

Previous studies have applied machine learning and deep learning techniques to intrusion detection and social-media security. Etuh et al. (2021) reviewed social media network attacks and their preventive mechanisms and identified the potential of intelligent intrusion detection and machine-learning approaches, while noting the limited availability of practical implementations and real-world evaluation. More recent studies have demonstrated the effectiveness of deep-learning architectures for cyberattack detection. For example, Al-Hadhrami et al. (2022) and Haq et al. (2022) applied hybrid LSTM-CNN and CNN-LSTM models to structured IoT intrusion datasets and reported accuracies above 99%. However, these studies focused primarily on network and IoT environments rather than the linguistic and behavioral characteristics of social-media content. Similarly, Susilo et al. (2025) combined autoencoders, LSTM and CNN with SMOTE and achieved a 99.19% F1-score on the CIC-IoT 2023 dataset, although generalization and computational overhead remained challenges. Within the social-media domain, Adekunle et al. (2024) developed the Social Engineering Assault Detection (SEAD) framework using NLP and RNN-LSTM to detect pretexting, phishing, scareware, clickbait and quid pro quo attacks from Facebook posts. The model achieved a precision of 0.85 and recall of 0.79. However, its

focus on Facebook and text-based data limits its applicability to other platforms and forms of social-media interaction. Rovito et al. (2022) also explored interpretable bot detection on Twitter using genetic algorithms and genetic programming, but reported a lower accuracy of approximately 75% and focused primarily on detection rather than prevention. Overall, the reviewed studies demonstrate the effectiveness of deep-learning techniques for cybersecurity but reveal gaps in cross-platform social-media detection, text-based social engineering detection, and the integration of complementary local and sequential linguistic features. The present study addresses these gaps by combining CNN with RNN-LSTM and evaluating the resulting hybrid model on balanced Facebook and Twitter social-media data. The proposed approach achieved an accuracy of 0.95, precision of 0.95, recall of 0.89, F1-score of 0.91 and AUC of 0.89

Materials and Methods

Dataset and Data Preparation

The experimental dataset contained 20,000 benign and malicious social media posts drawn from Facebook and Twitter. The malicious content represented social-engineering-related threats, including phishing, impersonation, clickbait and scareware, while benign samples represented legitimate social-media interactions. The dataset was prepared through cleaning, normalization and tokenization. The study also applied SMOTE to address class imbalance and produce a balanced dataset for model training and evaluation.

NLP and Hybrid Model Architecture

The model uses NLP-oriented preprocessing to convert raw textual interactions into a form suitable for neural-network learning. After preprocessing and representation, CNN layers extract local and n-gram-like linguistic patterns from the text. The resulting feature representations are passed to RNN-LSTM layers, which capture longer-range sequential dependencies and contextual information. The hybrid architecture is intended to combine complementary spatial/local and temporal/sequential information that may be useful for distinguishing deceptive from benign communication.

Model Training and Evaluation

The balanced dataset was divided into 80% training data and 20% testing data using stratified sampling. Categorical cross-entropy was used as the loss function and the Adam optimizer was used for model optimization. Batch normalization, dropout rates of 0.3–0.5 and early stopping based on validation loss were employed to reduce overfitting and improve generalization. The model classified inputs into benign and malicious categories. Accuracy, precision, recall, F1-score and ROC-AUC were used to assess performance.

Results

The proposed CNN-RNN-LSTM model achieved 95% overall accuracy on the unseen test set. Macro-averaged precision was 0.95, macro-averaged recall was 0.89 and macro-averaged F1-score was 0.91. The ROC-AUC was 0.89, indicating strong discrimination between benign and malicious samples. The training and validation curves showed progressive improvement with only minor divergence, which the dissertation interprets as evidence of stable convergence and limited overfitting.

Table 1. Overall Classification Performance of the Proposed Model

Metric	Score
Accuracy	0.95
Precision (macro)	0.95
Recall (macro)	0.89
F1-score (macro)	0.91
ROC-AUC	0.89

Table 2. Class-wise Performance

Class	Precision	Recall	F1-score
Benign	0.95	0.99	0.97
Attack (Malicious)	0.94	0.78	0.85

The confusion matrix further showed 3,199 true negatives and 36 false positives for the benign class, while 598 malicious samples were correctly identified and 167 malicious samples were classified as benign. This indicates that the model preserved legitimate activity effectively but still has room to improve malicious-class recall.

Discussion

The results support the usefulness of combining CNN and RNN-LSTM components for social-engineering detection in unstructured social-media text. The CNN component provides local pattern extraction, whereas the RNN-LSTM component models sequential dependencies. The dissertation reports that the hybrid model outperformed the standalone CNN and RNN-LSTM baselines by approximately 5–12% in accuracy. Comparison with the studies reported in the dissertation places the proposed model favorably within the social-media domain. Adekunle et al. (2024) reported precision of 0.85 and recall of 0.79 for an RNN-LSTM model applied to Facebook posts; the proposed model achieved macro precision of 0.95 and macro recall of 0.89. Other studies cited in the dissertation, including Al-Hadhrami et al. (2022) and Haq et al. (2022), reported higher accuracies on structured IoT datasets. Such results are not directly equivalent because the proposed study addresses the greater linguistic variability and semantic ambiguity of social-media text. Rovito et al. (2022), using an interpretable approach for Twitter bot detection, was reported at approximately 75% accuracy, while Alsubhi & Tawalbeh (2025) reported 98.88% in an attention-enhanced IoT intrusion setting.

Table 3. Comparison with Selected Studies Reported in the Dissertation

Study	Dataset/Domain	Model	Reported Result
Al-Hadhrami et al. (2022)	BoT-IoT	Hybrid LSTM-CNN	Accuracy 99.87%; Precision 99.89%; Recall 99.85%
Alsubhi & Tawalbeh (2025)	Standard IoT intrusion	CNN-LSTM with attention	Accuracy 98.88%
Haq et al. (2022)	UNSW-NB15 & X-IIoTID	Hybrid CNN-LSTM	Binary 99.84%; multi-class 99.80%
Rovito et al. (2022)	TwiBot-20 (Twitter)	Genetic Algorithms & Programming	Accuracy 75%
Adekunle et al. (2024)	Facebook posts	RNN-LSTM	Precision 0.85; Recall 0.79
Proposed study	Facebook & Twitter posts	CNN-RNN-LSTM	Accuracy 0.95; Precision 0.95; Recall 0.89; F1 0.91; AUC 0.89

The main practical implication is that a social-media security system must balance threat detection with preservation of legitimate user activity. The proposed model's high benign recall and relatively high precision reduce unnecessary alerts, while the attack recall of 0.78 at the class level shows a remaining limitation: some malicious posts are difficult to distinguish from legitimate communication. The dissertation therefore recommends continuous retraining with evolving datasets and future investigation of attention mechanisms, transfer learning and adversarial training.

Conclusion

This study developed and evaluated an Intelligent Intrusion Detection Model based on a CNN-RNN-LSTM hybrid architecture for social engineering attack detection on social media platforms. Using NLP-based preparation of textual data, SMOTE balancing and a cross-platform Facebook/Twitter dataset, the model achieved 95% accuracy, 0.95 macro precision, 0.89 macro recall, 0.91 macro F1-score and 0.89 AUC. The results demonstrate that combining local linguistic feature extraction with sequential modelling can provide an effective approach for detecting malicious social-media content. Although the model performed strongly overall, the lower malicious-class recall shows that further work is needed to capture subtle and evolving attacks. The study supports the use of hybrid deep-learning architectures as a practical foundation for proactive social-engineering detection on social media.

Acknowledgement

The author acknowledges the guidance and support received during the research and the institutional support provided by the Federal University Kashere.

Funding Information

This study was conducted with the support of a scholarship awarded by the Tertiary Education Trust Fund (TETFund), Nigeria, under its scholarship programme for my program (Master of Science (MSc) in Computer Science programme). The scholarship provided financial support throughout the study, including the research and related academic activities.

Conflict of interest

The authors declare that there is no conflict of interest, financial or otherwise, related to publication of this research paper.

Author Contributions

Sulaiman Ahmad conducted the research, including the conceptualization, data collection and preparation, development and evaluation of the proposed CNN-RNN-LSTM model, analysis and interpretation of results, and preparation of the manuscript. The research was carried out under the distinguished supervision and guidance of Professor P. B. Zirra, the author's amiable and greatly respected supervisor and academic mentor. His invaluable intellectual guidance, mentorship, constructive contributions, encouragement, and overall supervision played a significant role in the successful completion of this research.

Ethics

Not Applicable.

AI tool usage declaration

The authors declare that No AI-Assisted tools were used and No AI system was used to generate original scientific claims, data, or research results.

References

Abuali, K. M., Nissirat, L., & Al-Samawi, A. (2023). Intrusion detection techniques in social media cloud: Review and future directions. *Wireless Communications and Mobile Computing, 2023*, Article 6687023. https://doi.org/10.1155/2023/6687023

Adekunle, T. S., Lawrence, M. O., Alabi, O. O., Ebong, G. N., Ajiboye, G. O., & Bamisaye, T. A. (2024). The use of AI to analyze social media attacks for predictive analytics. *American Journal of Operations Management and Information Systems, 9*(1), 17–24.

Alenezi, M., & Eisa, M. A. (2022). Comparative evaluation of machine learning techniques for intrusion detection systems. *Computers & Security, 112*, 102510. https://doi.org/10.1016/j.cose.2022.102510.

Othman, A. M. H., Ab Razak, M. F., Firdaus, A., Tahaei, H., & Gheisari, M. (2026). A hybrid CNN–LSTM model for IoT intrusion detection: A robustness analysis across datasets. *Future Internet, 18*(7), 345.

Alsubhi, A., & Tawalbeh, L. (2025). *CNN-LSTM with attention for IoT intrusion detection*.

Etuh, E., Bakpo, F. S., & Agozie, E. H. (2021). Social media network attacks and their preventive mechanisms: A review. *Computer Science & Information Technology, 11*(4), 59–72. https://doi.org/10.5121/csit.2021.112405.

Altunay, H. C., & Albayrak, Z. (2023). A hybrid CNN + LSTM-based intrusion detection system for industrial IoT networks. *Engineering Science and Technology, an International Journal, 38*, 101322.

Kayes, I., & Iamnitchi, A. (2017). Privacy and security in online social networks: A survey. *Online Social Networks and Media, 3*, 1–21. https://doi.org/10.1016/j.osnem.2017.09.001.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. (2022). Text classification algorithms: A survey. *Information, 13*(1), 38.

Mitnick, K. D., & Simon, W. L. (2002). *The art of deception: Controlling the human element of security*. Wiley Publishing.

Rovito, L., Bonin, L., Manzoni, L., & De Lorenzo, A. (2022). An evolutionary computation approach for Twitter bot detection. *Applied Sciences, 12*(12), 5915.

Salahdine, F., & Kaabouch, N. (2019). Social engineering attacks: A survey. *Future Internet, 11*(4), 89. https://doi.org/10.3390/fi11040089.

Sarker, I. H., Ezugwu, A. E., & Chatterjee, J. M. (2021). Cybersecurity data science: An overview from machine learning perspective. *Journal of Big Data, 8*(1), 1–35.

Wang, Z., Ren, Y., Zhu, H., & Sun, L. (2022). Threat detection for general social engineering attack using machine learning techniques. *arXiv*. https://doi.org/10.48550/arXiv.2203.07933.

Yousef, A. M., Zain, J. M., & Alazab, M. (2023). SIEM-enhanced IPS with deep learning for high-speed networks. *Computers & Security, 122*, 102920. https://doi.org/10.1016/j.cose.2023.102920.

Zhou, Y., Qin, Z., Xie, S., & Wang, Y. (2023). A hybrid deep learning model for detecting cyberbullying on social media. *IEEE Transactions on Computational Social Systems, 10*(2), 345–356.