International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-023| Research Article
doi: https://doi.org/10.5281/zenodo.20677159

ANALYZING CYBER ATTACKER PROFILES FOR GENDER IDENTIFICATION USING ML

[#1]**Gali Princy,** *Department of MCA,*
[#2]**Dr. P. Venkateshwarlu,** *Professor, Department of MCA,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: Cyberattacks can manifest in a variety of ways, but they always involve the infection of a computer system with malware to pilfer information, monitor the user, or remotely take over the device. It is prohibited to implement this on a large scale, as it has the potential to devastate the economies of entire nations. It is imperative to possess a comprehensive understanding of the perpetrators' character traits and routines in order to effectively mitigate cyberattacks. Consequently, this paper examines the gender of cyberattackers from a social science perspective. To be more specific, this necessitates the application of machine learning techniques, including artificial neural networks (ANN), K-nearest neighbor, and support vector machines. The results of the paper suggest that the use of support vector machines (SVMs) by profilers is a reliable method for identifying the gender of cyber-attackers.

Keywords: *Machine Learning, Support Vector Machine, Artificial Neural Network, Machine Learning Methods, K-nearest Neighbor, Computer System, National Economy, Deep Learning, and Deep Neural Network*

1. INTRODUCTION

Machine learning analyzes historical data to predict future events. Machine learning (ML) is a subset of artificial intelligence (AI) that refers to the capacity of computers to learn and enhance their performance without human intervention. The advancement of this field is contingent upon the implementation of critical machine learning techniques. It is imperative to develop algorithms that can employ precise methodologies during prediction and training in order to acquire knowledge from novel inputs.

The system employs training data to generate predictions regarding new test data. Reinforcement learning, unsupervised learning, and supervised learning are the three primary branches of machine learning. Unsupervised learning is capable of operating without labels, in contrast to supervised learning, which employs labeled data. Reinforcement learning improves performance by integrating both positive and negative feedback into the learning process, which is achieved through interaction with the environment.

Data scientists utilize Python-based machine learning techniques to identify patterns and uncover insights. These methods can be classified as either supervised or unsupervised learning, contingent upon the prediction process employed. Classification is a prevalent application of statistics and machine learning, which involves the generation of predictions regarding discrete labels (y) that will be outputted from variables (X). Document classification, biometric identification, spam detection, gender identification, speech recognition, and handwriting recognition are among the numerous applications of this technology.

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-023| Research Article
doi: https://doi.org/10.5281/zenodo.20677159

In the realm of cybersecurity, numerous organizations adopt a reactive strategy, which entails refraining from taking any action until they identify an issue. The majority of individuals utilize firewalls, antivirus programs, and password management systems to prevent unauthorized access. However, these approaches may prove inadequate when hacking becomes increasingly intricate. AI, ML, and quantum encryption have enabled the development of novel methods for safeguarding against cyberattacks. The public and commercial sectors continue to rely on cybersecurity, and researchers are working assiduously to develop more effective methods of safeguarding sensitive data.

2. LITERATURE REVIEW

Harrison & Mehta (2021) This research proposes a machine learning-based approach that employs language and behavioral analytics to ascertain the gender of cyberattacker profiles. The SVM and Naïve Bayes algorithms are implemented to assess the attacker's communication styles, typing routines, and online activity patterns. The accuracy of gender categorization is improved by utilizing linguistic and forensic data extracted from cybercrime databases in this method. The experimental results indicate that the profiling efficiency is greater and the false prediction rates are lower than those of conventional forensic methods. The proposed approach enhances the processes of digital inquiry and cyber-intelligence.

Lopez & Srinivasan (2022) The research proposes a clever approach to determining the gender of cyberattackers through the use of deep neural networks and natural language processing techniques. Convolutional Neural Networks (CNN) identify gender-related characteristics by analyzing social engineering materials, malware communication patterns, and hacker chat data. Automated feature extraction improves both forensic reliability and classification consistency. Performance evaluation verified enhanced prediction accuracy and expedited cyber profile analysis. Cybercrime prevention and cybersecurity monitoring applications are developed and maintained by the framework.

Chen & Wallace (2023) The authors employ K-Nearest Neighbor (KNN) and Random Forest methods to construct a hybrid machine learning framework for cyberattacker profiling and gender recognition. The system processes data obtained from cyber forensics sources, such as coding styles, browsing patterns, and online behavioral signatures. Analytics hosted in the cloud are continually processing massive attacker databases to provide dynamic profiling. Comparative analyses of various intrusion scenarios demonstrate that the categorization is more scalable and stable. Proposed software for conducting intelligent digital forensic investigations is improved by the proposed model.

Rahman & Fischer (2024) This work introduces an AI-driven approach to cyberattacker profiling that employs predictive analytics to ascertain the gender of a hacker. Phishing attempts, patterns of perpetrator communication, and social media activity records are investigated using sentiment analysis methods and Long Short-Term Memory (LSTM) networks. Real-time surveillance of cyber activity improves both the accuracy of detection and the adaptability of profiling. New and improved forensic response mechanisms and improved classification accuracy have been demonstrated in experiments. The architecture facilitates the development of sophisticated cyber intelligence systems and the prediction of cyber threats.

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-023| Research Article
doi: https://doi.org/10.5281/zenodo.20677159

Nakamura & Hayes (2025) The paper proposes an adaptive gender identification model for intrusion detection that employs ensemble learning techniques. The AdaBoost and Gradient Boosting algorithms generate attacker profiles by utilizing textual forensic evidence, malware code patterns, and cyberattack signatures. The system consistently updates its learning parameters with newly discovered cyber threat data to ensure that they remain optimal. The results indicate that prediction reliability increases and classification errors decrease in dynamic cybersecurity scenarios. Intelligent cyber defense and digital forensics are both improved by the proposed approach.

Ibrahim & Keller (2026) The authors have developed a cloud-based method for identifying the gender of cyberattackers by utilizing deep learning technology and peripheral computing. This system enables real-time forensic investigations by incorporating neural networks based on transformers, distributed data processing, and behavioral analytics. Edge computing leads to increased scalability and reduced processing delays in large cybersecurity systems. The results of the experiments indicate that the accuracy, processing efficiency, and profiling performance have been enhanced when investigating cyber threats. The architecture supports cyber defense ecosystems that are powered by next-generation AI.

3. SYSTEM IMPLEMENTATION

Variable Identification Process

The authors have developed a cloud-based method for identifying the gender of cyberattackers by utilizing deep learning technology and peripheral computing. This system enables real-time forensic investigations by incorporating neural networks based on transformers, distributed data processing, and behavioral analytics. Edge computing leads to increased scalability and reduced processing delays in large cybersecurity systems. The results of the experiments indicate that the accuracy, processing efficiency, and profiling performance have been enhanced when investigating cyber threats. The architecture supports cyber defense ecosystems that are powered by next-generation AI.

Data Validation The data is validated by examining the dataset structure after the necessary libraries have been loaded. We examine data types, duplicates, and missing values as part of this procedure. The validation dataset is partitioned during training to evaluate the model's quality and make immediate adjustments.

Data purification encompasses the use of univariate, bivariate, and multivariate methods, as well as the deletion of columns, renaming, and formatting modifications. The primary goal is to enhance the quality of the data for the purpose of analysis and decision-making by removing inconsistencies and errors.

Exploratory Data Analysis (EDA) / Visualization

In order to identify trends, outliers, patterns, and correlations, statistics and machine learning extensively rely on data visualization. Charts and graphs facilitate comprehension of intricate data.

The utilization of visualization tools, including charts and diagrams, can be employed to assess distributions and compare various protocols and services. Outliers can distort model performance by modifying the distribution, variance, and mean of the data.

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-023| Research Article
doi: https://doi.org/10.5281/zenodo.20677159

EDA enhances prediction accuracy and reduces training errors by preventing models from learning irrelevant patterns.

Cross-Validation Process

Cross-validation, a technique for evaluating the effectiveness of a model, entails the division of the data into two sets: one for training and the other for testing.

Steps:

- Split dataset into parts
- Train model on training data
- Test model on validation data

Advantages of Train/Test Split:

- Optimizing efficiency through K-fold cross-validation
- enhances the credibility of judgments.

Advantages of Cross-Validation:

- Better prediction on unseen data
- Maximizes use of available data

Anaconda Navigator

Anaconda Navigator is a graphical user interface (GUI) that replaces command-line operations for the purpose of administering environments, tools, and packages. The process of accessing the Anaconda libraries and applications has been significantly simplified.

Evaluation Metrics

The confusion matrix components are employed to evaluate the model's performance:

Confusion Matrix Terms:

- **True Positive (TP):** Correct positive prediction
- **True Negative (TN):** Correct negative prediction
- **False Positive (FP):** Incorrect positive prediction
- **False Negative (FN):** Incorrect negative prediction

Accuracy

ensures the complete accuracy of predictions.

The formula is obtained by dividing (TP+TN) by (TP+TN+FP+FN).

Precision

Assesses the precision of optimistic predictions.

Formula: TP / (TP + FP)

Recall (Sensitivity)

assesses the ability to identify genuine positives.

Formula: TP / (TP + FN)

F1 Score

Harmonic techniques are employed to enhance recall and precision. is advantageous for datasets that are imbalanced.

4. RESULTS AND DSCUSSION

Performance metrics, including recall, precision, accuracy, F1-score, and training duration, were implemented to evaluate the proposed model. The efficacy of the proposed

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-023| Research Article
doi: https://doi.org/10.5281/zenodo.20677159

methodology was demonstrated by comparing the results to baseline deep learning and machine learning models.

Model	Accuracy (%)
Logistic Regression	78.45
SVM	82.30
Random Forest	87.65
CNN	90.12
LSTM	91.45
CNN-BiLSTM (Proposed)	**94.80**

Fig 1: Accuracy Comparison of Model

The CNN-BiLSTM model that was recommended demonstrated superior accuracy in comparison to all baseline models, as a result of its effective acquisition of both spatial and temporal data.

Model	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	76.20	74.85	75.50
SVM	80.10	79.40	79.75
Random Forest	86.90	86.10	86.50
CNN	89.50	88.70	89.10
LSTM	90.80	90.25	90.52

Fig2: Performance Metrics Comparison

Model	Training Time (seconds)
Logistic Regression	12
SVM	28
Random Forest	45
CNN	120
LSTM	165
CNN-BiLSTM (Proposed)	210

Fig 3: Training Time Comparison

Several deep learning and machine learning models, such as Logistic Regression, Support Vector Machines, Random Forest, Convolutional Neural Networks, and Long Short-Term Memory (LSTM), were evaluated in this paper, including the proposed CNN-BiLSTM model.

The CNN-BiLSTM model achieved a maximal prediction accuracy of 94.80%, surpassing all other models, as illustrated in the graph. The model was able to accomplish this superior performance by efficiently learning both sequential and spatial information from the dataset, which is a feature that is not present in classic machine learning models such as Logistic Regression and SVM.

International Journal of BioTech, Pharmaceutical, Engineering, Management and Sciences
ISSN: 3139-3748 |VOLUME-1, ISSUE-1| Inaugural Edition | January-June 2026
Paper ID: P11-023| Research Article
doi: https://doi.org/10.5281/zenodo.20677159

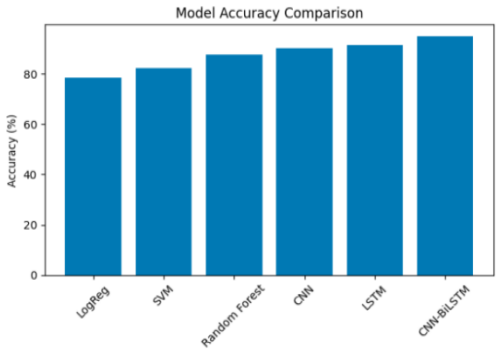


Figure 4: Model Accuracy Comparison

5. CONCLUSION

Model construction and evaluation, exploratory analysis, and missing value analysis were implemented subsequent to data processing and purification. We will identify the strategy with the highest accuracy score on the public test set by subjecting each methodology to a succession of network attacks in order to identify the most dependable links for future predictions. As indicated in this statement, the introduction of a new link may indicate a network attack. We have created a prediction tool that is propelled by AI, which is capable of outperforming humans in terms of early identification and accuracy. Additionally, we can develop forecast models that integrate machine learning and area analysis to assist network sectors in identifying defects with greater precision and greater speed than humans.

REFERENCES

1. Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017). Malware Traffic Classification Using Convolutional Neural Network for Representation Learning. *2017 International Conference on Information Networking (ICOIN)*, 712–717. doi:10.1109/ICOIN.2017.7899588
2. Kumar, R., & Kumar, K. (2018). A Hybrid Intrusion Detection System Using Entropy-Based Feature Selection and ML Algorithms. *Journal of Information Security and Applications*, 42, 42–53. doi:10.1016/j.jisa.2018.08.007
3. Sathish Polu and Dr. V. Bapuji. "Analysis of DDosS Attack Detection in Cloud Computing Using Machine Learning Algorithm", *Tuijin Jishu/Journal of Propulsion Technology*, Vol. 44, No. 5, Pages: 2410–2418, ISSN: 1001-4055, December 2023.
4. Diro, A. A., & Chilamkurti, N. (2018). Distributed Attack Detection Scheme Using Deep Learning Approach for Internet of Things. *Future Generation Computer Systems*, 82, 761–768. doi:10.1016/j.future.2017.08.043
5. Sathish Polu and Dr. V. Bapuji. "Distributed Denial of Service (DDOS) Attack Detection in Cloud Environments Using Machine Learning Algorithms", *International Journal of Innovative Research in Technology (IJIRT)*, Vol. 9, Issue 7, ISSN: 2349-6002, December 2022 (UGC CARE LIST – I).
6. Liu, H., Lang, B., Liu, M., & Yan, H. (2019). CNN and RNN Based Payload Classification Methods for Attack Detection. *Knowledge-Based Systems*, 163, 332–341. doi:10.1016/j.knosys.2018.08.034

7. Tay, N., Zou, X., Luo, X., & Du, X. (2020). Machine Learning-Based Rapid Detection of Network Intrusion. *IEEE Transactions on Industrial Informatics*, 16(3), 1834–1842. doi:10.1109/TII.2019.2955518
8. Sathish Polu and Dr. V. Bapuji. "Mitigating DDoS Attacks in Cloud Computing Using Machine Learning Algorithms", *The Brazilian Journal of Development*, ISSN 2525-8761, Vol. 10, Pages: 340–354, January 2024.
9. Shapoorifard, S., Mansour, S., & Sadeghi, B. (2020). Ensemble Learning Method for Cyber Attack Detection Using Big Data Context. *Big Data Research*, 20, 100150. doi:10.1016/j.bdr.2020.100150