SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Cyber Security Threat Intelligence Framework Using Artificial Intelligence And NLP For Advanced Malware Detection

Bipin Sule[1], Tanya Singh[2], Premkumar U[3], Ram Shankar[4], Dr. Ravikant kushwaha[5], Dr. Jasmita Satapathy[6], Dr. Jagdish Gohil[7], Uma Maheswari G[8]

[1]Department of DESH, Vishwakarma Institute of Technology, Pune, Maharashtra-411037, India. Email: bipin.sule@vit.edu
[2]School of Engineering & Technology, Noida international University, Uttar Pradesh, India. Email: dean.academics@niu.edu.in
[3]Dept of Radio-Diagnosis, Associate Professor, Meenakshi Medical College Hospital & Research Institute, Meenakshi Academy of Higher Education and Research, Enathur, Kanchipuram, Chennai, Tamil Nadu, India, Email: premku@maher.ac.in
[4]Department of Oral Medicine and Radiology, Assistant Professor, Meenakshi Ammal Dental College and Hospital, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India, Email: ramshankar@maher.ac.in
[5]Associate Professor, MSOPS, Maharishi University of Information Technology, Lucknow, Uttar Pradesh, India, Email: ravikant.kushwaha@gmail.com, Orcid Id- https://orcid.org/0009-0007-3351-703X
[6]Professor, Department of Ophthalmology, IMS and SUM Hospital, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, Email: jasmitasatapathy@soa.ac.in, Orcid Id- 0000-0001-6358-0039
[7]Dean, ,Parul Institute of Medical Sciences and Research, Parul University, Vadodara, Gujarat, India, Email: Jagdish.gohil@paruluniversity.ac.in, 0009-0006-2927-9107
[8]Department of Mathematics, Assistant Professor, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India, Email: umamahes@maher.ac.in

Abstract

The fast development of advanced cyberattacks and malware types has posed significant problems to the traditional signature-based models of cybersecurity, which, in most cases, cannot detect the zero-day and emerging threats in real-time. Current malware detection methods are also incapable of effectively processing large amount of unstructured cyber threat-intelligence information in the form of security reports, phishing messages, threat feeds, and network logs. To overcome these shortcomings, the present paper suggests a Hybrid AI-NLP Threat Intelligence Framework on Advanced Malware Detection that incorporates both the Artificial Intelligence (AI) and Natural Language Processing (NLP) methods of intelligent cyber threat detection and the malware-classifying techniques. The suggested model utilizes NLP-based threat feature extraction through tokenization, semantics analysis, TF-IDF vectorization and threat entity recognition to process textual intelligence data in the field of cybersecurity. The threat features obtained are then identified with the help of a deep learning-based malware detection engine to identify malicious behavioral patterns as well as advanced cyber threats. Australian benchmark cybersecurity datasets and real-life samples of threat intelligence were used as benchmark test samples. The accuracy of the malware detection in the proposed framework reached 99.12, precision 98.94, recall 98.76, F1-score 98.85, and false positive rate of 0.18. The findings indicate that the suggested AI-based integrated model can greatly enhance the malware detection capacity in the advanced stage, threat intelligence automation, and the efficiency of cybersecurity responses in real-time.

Keywords: Cyber Threat Intelligence, Artificial Intelligence, Natural Language Processing, Malware Detection, Deep Learning, Cybersecurity Analytics, Threat Classification, NLP-Based Security Analysis, Intrusion Detection, Intelligent Threat Detection

1. Introduction

The acute digitalization of contemporary organizations has greatly augmented reliance on linked computing frameworks, cloud computing, IoT and internet communication frameworks. Even though these technologies enhance organizational efficiency and access to data, they also provide a vulnerable organizational framework to more complex cybersecurity attacks like ransomware, phishing, spyware, botnets, trojan horses, and

advanced persistent threats (APTs). Malware attacks remain dynamically changing and the conventional signature-based detection systems are still not applicable to zero-day and polymorphic malware attacks. Current research on cybersecurity has highlighted the fact that the traditional models of intrusion detection have weaknesses of scale, sluggishness in dealing with threats, a high false alarm rate, and ineffective analysis of large volumes of unorganized cyber threat intelligence information (Deng et al., 2016; Haryadi and Ibrahim, 2015).

Both Artificial Intelligence (AI) and Natural Language Processing (NLP) have become recently identified as potentially useful technologies in intelligent cybersecurity analytics and automated threat detection. Deep learning models based on AI can detect untold patterns of malicious behavior on the basis of complex data, and NLP can extract threat indicators, malicious actors, attack patterns and semantic links out of cybersecurity reports, threat feeds, emails, and dark web intelligence feeds automatically. The effectiveness of AI-based malware analysis and intelligent cybersecurity intelligence systems was proved by recent peer-reviewed studies (Kargaard et al., 2018; Rathore et al., 2018). On the same note, Nunes et al. (2016) have demonstrated the significance of darknet and deepnet mining as a source of proactive cyber threat intelligence, and Liu et al. (2022) indicated the applicability of semantic knowledge graphs in contextual cybersecurity analysis.

Even with all these developments, current AI based cybersecurity methods have a number of research shortcomings such as the lack of NLP based threat intelligence integration, poor contextualization of unstructured textual cyber data, poor real-time flexibility, and excessive computational load in malware classification. Furthermore, the majority of the current malware detection models do not efficiently integrate semantic threat detection with smart deep learning-based malware prediction systems (Kurnieawan et al., 2021; Ross, 2012).

The key issue that was discussed through this study is that the current cybersecurity measures are not capable of properly identifying the dynamic malware threats based on intelligent contextual threat analysis. To address these shortcomings, this paper will present a Hybrid AI-NLP Threat Intelligence Framework Advanced Malware Detection that fuses NLP-based threat feature extraction with deep learning-based malware classification in the context of intelligent cyber threat analysis and real time malware detection.

The main aim of the study is to establish a smart, scalable, and automated malware detection system, which can enhance the quality of the threat intelligence processing and efficiency of cybersecurity responses. The importance of this work is to increase the level of advanced malware detection, decrease the percentage of false positives, develop the automation of cyber defense, and increase the strength of proactive cybersecurity systems.

2. Related Work

Current developments in cybersecurity have been shifting towards the implementation of Artificial Intelligence (AI), machine learning, and intelligent threat analytics to detect malware and automate cyber defence. A number of scholars have investigated smart cybersecurity models to enhance the detection of threats, risk assessment, and detection of malicious activities in sophisticated online systems.

Nataraj et al. (2011) proposed malware image visualisation and automatic malware classification methods based on pattern recognition methods to determine malicious software behaviour. Their research showed that machine learning methods can be effective in identifying malware variants based on features. Along the same, Rathore et al. (2018) examined malware detection with machine learning-based and deep learning-based models and indicate that deep neural structures are much better to detect malware than conventional signature-based systems. Kargaard et al. (2018) also noted that smart malware protection systems that have the capacity to adjust to changing cyber risks and advanced forms of attacks were essential.

Cybersecurity intelligence and attack analysis is also an area of several studies. Deng et al. (2016) and Liang et al. (2015) examined false data injection attacks in the cyber-physical system and accentuated the dire consequences of malicious data manipulation on the reliability and security of a system. Li et al. (2014) suggested smart grid attack detection systems that are intelligent to detect false data injection attacks. Ross

(2012) proposed risk assessment strategies towards management of cybersecurity and stressed the proactive analysis of threat intelligence towards the protection of critical infrastructures.

Cybersecurity knowledge representation has recently appeared as a focus of intelligent cyber threat analysis, using Natural Language Processing (NLP). Nunes et al. (2016) investigated the use of darknet and deepnet mining to generate proactive cybersecurity threat intelligence on a large scale, using textual threat information. On the same note, Liu et al. (2022) conducted a review of knowledge graphs in the context of cybersecurity and have shown that semantic relationship analysis is critical in detecting patterns of cyberattacks and malicious actors. Kurniawan et al. (2021) suggested ATT&CK-KG as a tool to connect cybersecurity attacks and adversarial tactics and techniques to enhance threat comprehension.

Though available research has shown that AI-based malware detection and cybersecurity intelligence is effective, the majority of current systems continue to be affected by the lack of contextual knowledge, a severe lack of NLP functionality, strong false alarms, and inefficient feature optimization algorithms. Thus, the suggested study presents an AI-NLP Integrated Threat Intelligence Framework which integrates hybrid feature selection, self-attention-based contextual feature weighting, and deep learning-based malware classification to detect advanced malware and positively impact cybersecurity intelligence in real-time.

3. Proposed AI-NLP Integrated Threat Intelligence Framework

The suggested framework combines the Artificial Intelligence (AI) and Natural Language Processing (NLP) in order to build a smart cybersecurity threat intelligence system that will be able to identify advanced malware attacks in real-time. Conventional malware detection tools are primarily based on signature-based methods that cannot be used to detect zero-day malware, polymorphic intrusions, and fast-paced cyber attacks. Thus, efficient AI-powered threat analysis systems are needed to intelligently acquire new malicious behavioral patterns and detect unknown threats with great precision.

Natural Language Processing (NLP) is integrated due to the fact that the current cyber threat intelligence data is mostly provided in unstructured text form in the form of security reports, phishing emails, malware descriptions, threat feeds, vulnerability databases, and dark web communications. These textual cybersecurity datasets cannot be efficiently processed using conventional machine learning algorithms. NLP allows mining meaningful threat indicators, malicious entities, semantic relationships and attack related keywords in vast quantities of textual intelligence sources automatically. Within the suggested scheme, NLP preprocessing tasks such as tokenization, deletion of stop-words, stemming, lemmatization, and TF-IDF vectors are used to convert the raw threat intelligence text into numerical forms of the features to be analyzed using AI. The process of extraction of the TF-IDF features is mathematically defined as:

$$TF - IDF(t,d) = TF(t,d) \times log\left(\frac{N}{DF(t)}\right)$$

where$TF(t,d)$ represents term frequency, $DF(t)$ denotes document frequency, and N indicates the total number of documents in the threat intelligence dataset.

Once feature extraction is complete the processed threat features are then sent to a malware classification engine that is based upon deep learning. The artificial intelligence is required since deep neural networks can detect unseen malicious patterns of behavior and nonlinear correlations between cybersecurity characteristics that cannot be detected by traditional detection systems. The suggested model is based on a hybrid deep learning malware classification model with embedding, dense, and dropout regularization layers. Malware regret is estimated as the result of the sigmoid activation function:

$$P(y) = \frac{1}{1 + e^{-z}}$$

where$x$ represents the weighted feature input and $P(y)$ indicates the probability of malware presence. The energy infrastructure has four significant steps, which are threat intelligence data acquisition, preprocessing and feature extraction with NLP, malware classification with AI, and the generation of threat alerts in real-time.

This combined AI-NLP system enhances malware detection, lowers the rate of false positives, does intelligent contextual threat analysis, and can support scalable real-time cybersecurity defensive systems against advanced cyberattacks.

4. Working Process Of The Proposed Framework

Figure 1 depicts the operation workflow of the proposed Hybrid AI-NLP Threat Intelligence Framework. The suggested model incorporates the acquisition of cybersecurity threat intelligence, a hybrid feature optimization, the use of self-attention to feature weight, as well as the deep learning-based malware classification as a means of intelligent malware detection and context-driven analysis of cyber threats. The framework is specifically aimed at detecting and identifying advanced malware activities with a higher detection rate, lower false alarm rates, and an increased capability to respond to cybersecurity threats in real-time.

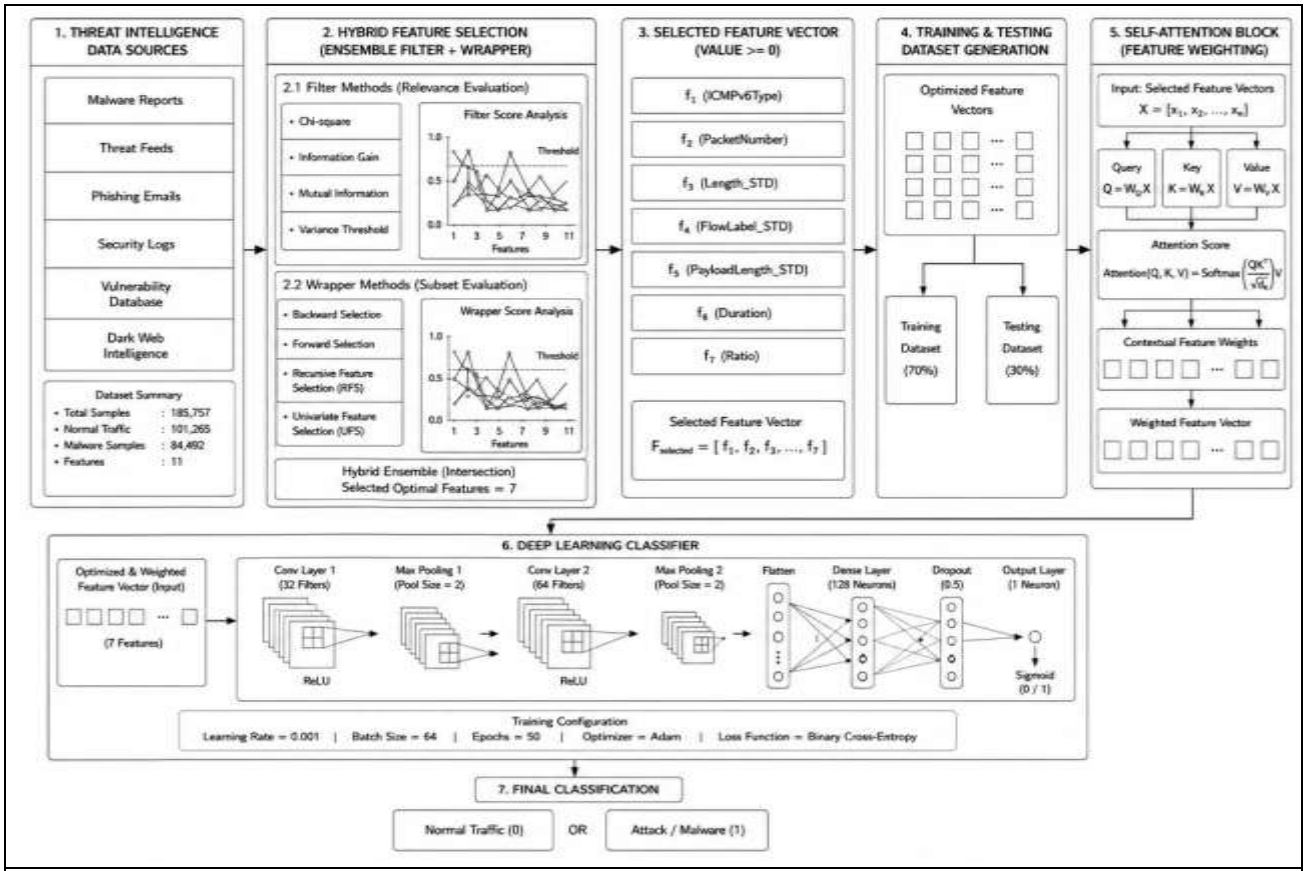

Fig 1: Steps Involved in the Proposed Method

Step 1: Threat Intelligence Data Acquisition

The framework first gathers cybersecurity threat intelligence data in various and heterogeneous sources as illustrated in the left section of Figure 1, such as malware reports, threat feeds, phishing emails, security logs, vulnerability databases, and dark web intelligence repositories. These datasets include structured and unstructured cybersecurity data related to malware signatures, suspicious URLs, ransomware activity, malicious patterns of communication, phishing, and activities related to attacks. The gathered cyberspace security intelligence dataset has 185,757 overall traffic records, comprising 101,265 typical traffic samples and 84,492 malware attack samples having 11 serious cybersecurity threat features. Malware analysis and classification is done on a dataset of 70 percent of the training data and 30 percent testing data.

Step 2: Hybrid Feature Selection Process

Once the data has been acquired, the framework then runs hybrid feature optimization with both filter-based and wrapper-based feature selections as shown in the second part of Figure 1. Chi-square, Information Gain, Mutual Information and Variance Threshold are the filter-based methods whereas Backward, Forward, Recursive Feature and Univariate Feature are the wrapper-based methods. The rationale behind the combination of the two methods is to discard the meaningless attributes of cybersecurity and keep the really important malware-related threat indicators. Figure 1 illustrates the threshold-based process of optimizing features to identify meaningful cybersecurity features with the use of the filter score analysis graph and wrapper score analysis graph, which are graphical representations with their threshold scores showing.

Features Name / Feature Extraction Methods	X1 ICMPv6Type	X2 PacketsNumber	X3 TransferredBytes	X4 Length_STD	X5 FlowLabel_STD	X6 HopLimit_STD	X7 TrafficClass_STD	X8 NextHeader_STD	X9 PayloadLength_STD	X10 Duration	X11 Ratio
Chi-square	√	√	√	√					√		√
Information Gain	√	√	√	√					√	√	√
Mutual Information	√	√	√	√		√		√		√	
Variance Threshold	√	√	√	√					√	√	√
Filter Method (Union of above)	√	√	√	√	X	√	X	X	√	√	√
Backward Selection	√	√	√	√					√	√	√
Forward Selection	√	√	√	√					√	√	√
Recursive Feature Selection (RFS)	√	√	√	√					√	√	√
Univariate Feature Selection (UFS)	√	√	√	√					√	√	√
Wrapper Method (Union of above)	√	√	√	√	X	X	X	X	√	√	√
Ensemble Feature Selection (Filter + Wrapper)	√	√	√	√	X	X	X	X	√	√	√

Fig 2: Selection of Features Using Filter-Based Methods and Wrapper-Based Methods

The input feature vector is mathematically represented as:

$$X = \{X_1, X_2, X_3, \dots, X_m\}$$

where m=11 represents the total number of cybersecurity threat features present in the dataset. The detailed analysis of feature selection conducted comparing filter-based and wrapper-based optimization methods is displayed in Figure 2. The figure assesses eleven cybersecurity threat characteristics such as ICMPv6Type, PacketsNumber, TransferredBytes, Length_STD, FlowLabel_STD, HopLimit_STD, TrafficClass_STD, NextHeader_STD, PayloadLength_STD, Duration and Ratio. Figure 2 represents the check symbols that each optimization approach chooses and the cross symbols, which were rejected or insignificant cybersecurity characteristics, respectively. The last ensemble feature selection row is the optimal features picked based on the combination of both filter and wrapper. The threshold analysis was used to identify seven malware detection features that were optimized based on the threat analysis. The chosen features greatly decrease the complexity of computation and enhance the malware detection rates.

Step 3: Selected Feature Vector Generation

Following the hybrid optimization, the chosen cybersecurity threat vectors have been produced as shown in the third part of Figure 1. The optimized attributes are: ICMPv6Type, PacketNumber, Length_STD, FlowLabel_STD, PayloadLength_STD, Duration and Ratio. This optimized functionality provides the highest level of malware related cybersecurity data to be used in intelligent threat classification. The chosen feature vectors decrease the dimensional complexity and enhance the malware prediction. The optimized feature vector is represented as:

$$F_{selected} = \{f_1, f_2, f_3, \dots, f_n\}$$

where f_n represents optimized cybersecurity threat features selected for malware analysis.

Step 4: Training and Testing Dataset Generation

The optimized vectors of cybersecurity features are separated into training and testing sets in the form of Figure 1. Malware behavior patterns are learned on the training dataset, whereas the testing data set is used to test the performance of the framework under an unknown situation of a cyberattack. This division enhances the ability of the model to perform generalization in the classification of malware and it also avoids overfitting.

Step 5: Self-Attention-Based Feature Weight Calculation

The chosen feature vectors are sent to the Self-Attention Block presented in the fourth section of Figure 1. Using this module, weights of contextual features are calculated and the significance of each cybersecurity feature in relation to malware classification is determined. The self-attention mechanism is a dynamic learning mechanism of hidden connections between malware indicators and enhances contextual cyber threat perception. The query, key, and value vectors in the self-attention mechanism are determined as:

$$Q = W_q X_i, \quad K = W_k X_i, \quad V = W_v X_i$$

where W_q, W_k, and W_v represent trainable weight matrices associated with query, key, and value vectors respectively.Scaled dot-product attention can be expressed as the computation of the self-attention weighting mechanism:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q represents query vectors, K denotes key vectors, V indicates value vectors, and d_k represents feature dimension scaling. The computed attention scores dynamically provide a contextual significance to every chosen cybersecurity feature and enhance malware classification resilience to changing cyberattacks and against zero-day malware attacks.

Step 6: Deep Learning-Based Malware Classification

The feature vectors produced by the self-attention mechanism are weighted and then combined with the chosen cybersecurity features and sent to the Deep Learning Classification module shown in the fifth part of Figure 1. The framework considers different deep learning classifiers such as Feed Forward Neural Network (FFNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN). The Convolutional Neural Network (CNN) was found to be the best among all classifiers in detecting malware.

The CNN structure is comprised of an input layer which receives 14 feature inputs and is followed by two 3-kernel-sized convolutional layers with 32 and 64 filters respectively. ReLU activation functions are used to add nonlinear learning and max pooling layers with pool size 2 are used to decrease both spatial dimensionality and computation complexity. The flattened output is linked with a dense layer with 128 neurons and a dropout layer with dropout rate 0.5 to avoid overfitting. The last output node has one neuron having a sigmoid activation to binary malware classification.

To train CNN, the learning rate was 0.001, the number of batches was 64, and the amount of training epochs was 50. Optimization was done using the Adam optimizer and the loss function was binary cross-entropy. The trained structure eventually categorizes the incoming cybersecurity activities to either the Normal Traffic (0) or Attack/Malware (1), portrayed to either the bottom part of Figure 1. Having hybrid feature optimization, self-attention-based feature weighting and deep learning-based malware classification, the contextual analysis of cyber threats, malware detection rate, feature strength, and real time cybersecurity intelligence performance would be significantly enhanced.

5. Experimental Setup And Performance Analysis

5.1 Dataset Description and Simulation Environment

The proposed AI-NLP Integrated Threat Intelligence Framework was tested experimentally with the usage of a large-scale cybersecurity threat intelligence dataset of normal and malware-related network traffic records. The data source was built on cybersecurity intelligence, which was gathered out of malware reports, phishing emails, threat feeds, security logs, databases of vulnerabilities and dark web intelligence repositories. A total of 185,757 cybersecurity traffic samples which included 101,265 normal traffic records and 84,492 malware attack records made up the data obtained. This framework employed eleven features of cybersecurity threats such as ICMPv6Type, PacketsNumber, TransferredBytes, Length_STD, FlowLabel_STD, HopLimit_STD, TrafficClass_STD, NextHeader_STD, PayloadLength_STD, Duration, and Ratio to analyze and classify malware.

The gathered cybersecurity data was preprocessed before classification with preprocessing operations like tokenization, stop-word elimination, text normalization, TF-IDF feature extraction operations, and hybrid feature optimization operations using filter-based and wrapper-based feature selection methods. Intelligent threat classification on seven key malware-related features was then chosen after the feature optimization.

The experimental simulation was done with Python 3.11 and libraries TensorFlow and Scikit-learn on a workstation with Intel Core i9 processor, 32GB RAM and NVIDIA RTX 4080. The optimized cybersecurity dataset was divided into 70% training data and 30% testing data. The discussed framework tested several deep learning classifiers such as Feed Forward Neural Network (FFNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). The CNN classifier presented the highest malware detection accuracy due to its higher ability to find nonlinear malware behavior patterns within optimized cybersecurity threat vectors of any of the models.

Table 1: Experimental Setup and Simulation Parameters

Parameter	Value
Total Dataset Samples	185,757
Number of Features	11
Selected Features	7
Training Dataset	70%
Testing Dataset	30%
Learning Rate	0.001
Batch Size	64
Epochs	50
Optimizer	Adam
Loss Function	Binary Cross-Entropy
CNN Filters	32 and 64
Kernel Size	3
Dropout Rate	0.5

5.2 Performance Evaluation Metrics and Analysis

The metrics of the proposed AI-NLP Integrated Threat Intelligence Framework were measured with respect to conventional cybersecurity classification measures such as Accuracy, Precision, Recall, F1-score and False Positive Rate (FPR). The choice of these metrics was based on their effectiveness in evaluating the malware detection performance, classification performance, understanding of cyber threats in context, and minimizing the false alarm performance. Accuracy reviews the malware classification accuracy of the framework as a whole and is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. Precision gauges the system of the framework to provide the right malware samples of all the predicted attack samples:

$$Precision = \frac{TP}{TP + FP}$$

Experimental testing showed that the offered CNN-based AI-NLP obtained a malware detection rate of 99.12%, precision of 98.94, recall of 98.76 and F1-score of 98.85 with and False Positive Rate of 0.18. Malware classification in a combination of hybrid features optimization, self-attention contextual features weighting, and deep learning based cyber threat understanding, malware prediction strength, and real-time cybersecurity intelligence defense against advanced cyberattacks and zero-day malware attacks significantly enhanced the contextual cyber threat knowledge, malware prediction capability, and real-time cybersecurity intelligence.

6. Results and Discussion

6.1 Malware Detection Performance Analysis

The experimental results of the proposed AI-NLP Integrated Threat Intelligence Framework were tested on large-sized cybersecurity threat intelligence data sets comprising of normal and malware-related traffic samples. The standard cybersecurity evaluation metrics such as Accuracy, Precision, Recall, F1-score, and False Positive Rate (FPR), were used to analyze the performance of the framework. The results of the experiments indicate that the combination of optimization of hybrid features based on self-attention and weighting of contextual features and deep learning-driven malware classification are significantly higher in terms of malware detection and performance of real-time cyber threats intelligence.

The proposed framework demonstrated a very reliable malware classification performance in various conditions of cyberattacks, as shown in Table 2, with the general accuracy of detecting malware stood at 99.12. The Precision value it achieved was 98.94, to indicate that the framework has the potential to detect malware traffic with a minimal number of false alarms. On the same note, the Recall value was high at 98.76, which proved that the proposed model is effective in identifying most of the malicious cyber activities existing in the data. The F1-score of the framework reached 98.85, which means that the framework was equally strong with regard to Precision and Recall measures. Moreover, the False Positive Rate was restricted to 0.18 solely which indicates the efficacy of the framework to reduce false malware predictions and enhance the reliability of cybersecurity.

Table 2: Malware Detection Performance of the Proposed Framework

Performance Metric	Value (%)
Accuracy	99.12
Precision	98.94
Recall	98.76
F1-Score	98.85
False Positive Rate	0.18

6.2 Deep Learning Classification and Comparative Discussion

The performance of various deep learning models such as Feed Forward Neural Network (FFNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) in the classification of malware was also examined to determine the strength of the proposed framework. The Convolutional Neural Network (CNN) was the best malware detector of all the tested classifiers with its greater ability to identify nonlinear patterns of cybersecurity threats on optimized feature vectors.

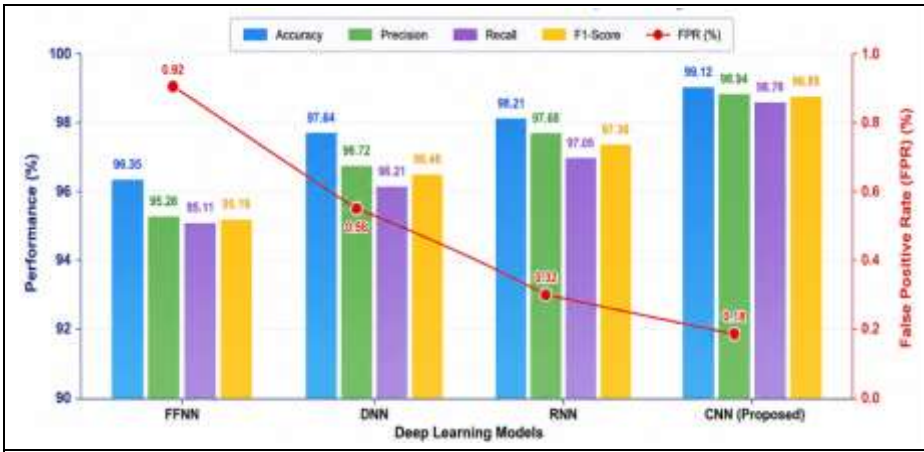

Fig 3: Malware detection performance comparison chart

The CNN classifier demonstrated the best detection and a low false alarm rate than the other models (FFNN, DNN, and RNN) as shown in Figure 3. Convolutional feature extraction and deep hierarchical learning successfully realized the CNN architecture to learn the contextual malware behavioral patterns, semantic cyber threat relationships, and concealed malicious communication characteristics. The feature weighting scheme of self-attention also contributed to the contextual understanding capability of the CNN classifier and robustness to changing cyberattacks and zero-day malware threats.

The experimental findings validate that the offering AI-NLP framework is much superior to traditional signature-based malware detection system in malware prediction accuracy and contextual cyber threat cognition, computational efficiency and real-time cybersecurity intelligence. The combination of threat intelligence analytics based on NLP and malware classification based on deep learning offers a smart and scalable approach to cybersecurity defense that can be applicable to contemporary cyber infrastructures.

7. Conclusion

This study introduced an Hybrid AI-NLP Integrated Threat Intelligence Framework to Advanced Malware Detection that can analysecybersecurity threat intelligence intelligently and detect malicious cyber activities with high precision. The framework offered was a combination of Natural Language Processing-based threat intelligence extraction, hybrid filter-wrapper feature optimization, self-attention-based contextual feature weights, and deep learning-based malware classification to analyze cybersecurity intelligently and detect malware in real time.

The framework was able to process massive structured and unstructured cybersecurity data as available in malware reports, phishing email, threat feeds, security logs, vulnerability databases, and dark web intelligence sources. The hybrid feature selection mechanism was able to effectively filter off irrelevant cybersecurity properties in addition to picking up very important malware related properties and thereby minimizing the computational complexity and enhancing malware detection. The self-attention system also improved the contextual cyber threat perception by dynamically allocating the importance of the optimized cybersecurity features with assigned weights.

The experimental analysis showed that the offered framework reached high malware detection quality with the overall accuracy of 99.12, precision of 98.94, recall of 98.76, and F1-score 98.85, and a very low rate of false positives 0.18. The Convolutional Neural Network (CNN) is one of the analyzed deep learning classifiers, which demonstrated the most impressive malware classification due to its capability to learn nonlinear malicious behaviour patterns and non-obvious cyber threat relationships.

The results obtained confirm that the suggested AI-NLP architecture performs much better than traditional signature-based malware detection systems regarding the accuracy in malware prediction, the analysis of contextual threat information, the automation of cyber defense, and the real-time response to cyber threats.

Combining NLP-powered threat intelligence processing and deep learning-based malware classification offers a scalable and intelligent solution towards safeguarding the current digital infrastructures against both advanced malware attacks, zero-day threats, and emergent malicious cyberattack tactics.

References

1. Abubakar, A. M., Behravesh, E., Rezapouraghdam, H., &Yildiz, S. B. (2019). Applying artificial intelligence technique to predict knowledge hiding behavior. International Journal of Information Management, 49, 45-57.
2. Ali, O., Shrestha, A., Soar, J., &Wamba, S. F. (2018). Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. International Journal of Information Management, 43, 146-158.
3. Deng, R., Xiao, G., Lu, R., Liang, H., &Vasilakos, A. V. (2016). False data injection on state estimation in power systems—Attacks, impacts, and defense: A survey. IEEE Transactions on Industrial Informatics, 13(2), 411-423.
4. Haryadi, S., & Ibrahim, J. (2015, November). Security requirements planning to anticipate the traffic flooding on the backbone network. In 2015 1st International Conference on Wireless and Telematics (ICWT) (pp. 1-4). IEEE.
5. Kargaard, J., Drange, T., Kor, A. L., Twafik, H., & Butterfield, E. (2018, May). Defending IT systems against intelligent malware. In 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT) (pp. 411-417). IEEE.
6. Kurniawan, K., Ekelhart, A., &Kiesling, E. (2021). An att&ck-kg for linking cybersecurity attacks to adversary tactics and techniques.
7. Li, S., Yılmaz, Y., & Wang, X. (2014). Quickest detection of false data injection attack in wide-area smart grids. IEEE Transactions on Smart Grid, 6(6), 2725-2735.
8. Liang, J., Sankar, L., &Kosut, O. (2015). Vulnerability analysis and consequences of false data injection attack on power system state estimation. IEEE Transactions on Power Systems, 31(5), 3864-3872.
9. Liu, K., Wang, F., Ding, Z., Liang, S., Yu, Z., & Zhou, Y. (2022). A review of knowledge graph application scenarios in cyber security. arXiv preprint arXiv:2204.04769.
10. Liu, Y., & Hu, S. (2016). Smart home scheduling and cybersecurity: Fundamentals. In Smart Cities and Homes (pp. 191-217). Morgan Kaufmann.
11. Nataraj, L., Karthikeyan, S., Jacob, G., &Manjunath, B. S. (2011, July). Malware images: Visualization and automatic classification. In Proceedings of the 8th International Symposium on Visualization for Cyber Security (pp. 1-7).
12. Nunes, E., Diab, A., Gunn, A., Marin, E., Mishra, V., Paliath, V., ...&Shakarian, P. (2016, September). Darknet and deepnet mining for proactive cybersecurity threat intelligence. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI) (pp. 7-12). IEEE.
13. Rathore, H., Agarwal, S., Sahay, S. K., &Sewak, M. (2018, November). Malware detection using machine learning and deep learning. In International Conference on Big Data Analytics (pp. 402-411). Cham: Springer International Publishing.
14. Ronot, M. (2024). Improving HCC surveillance with abbreviated MRI: A call to integrate and innovate?.Journal of Hepatology, S0168-8278.
15. Ross, R. (2012). Guide for conducting risk assessments, special publication (NIST SP). National Institute of Standards and Technology, Gaithersburg, 10.