Detecting Crypto Ransomware in Encrypted File Sharing Networks

Konda Srikar Goud(✉), Jaggaiahgari Roopa Sahithi, Sravya Vemula, and Barupati Harshitha

Department of Information Technology, BVRIT HYDERABAD College of Engineering for Women, Hyderabad, Telangana, India
Srikargoud.k@bvrithyderabad.edi.in

Abstract. Ransomware constitutes a form of malicious software designed to impede users' access to their systems, either by locking the system's screen or restricting access to files until a ransom is paid. In corporate settings, user computers typically store system and program files, with document access occurring through shared servers. In such instances, a single crypto-ransomware infected host can potentially block access to all shared files within its reach, encompassing the entire file set of a user workgroup. To address this issue, we propose implementing a machine-learning model to identify crypto-ransomware activity by analyzing file-sharing traffic. This framework actively monitors the traffic exchanged between clients and file servers, employing machine learning techniques to discern patterns indicative of ransomware actions, particularly during the reading and overwriting of files. Notably, the model is designed to operate not only for clear text protocols but also for encrypted file-sharing protocols. The model's efficacy is validated through an extensive dataset comprising over 70 ransomware binaries from 33 strains and more than 2,400 h of 'uninfected' traffic from actual users.

Keywords: Crypto-Ransomware · Clear text protocols · File sharing protocols · Ransomware binary · Machine Learning

1 Introduction

Crypto-ransomware poses a significant cyber security threat, exploiting encryption to hold users' data hostage and demanding ransom payments for recovery. In 2020, EUROPOL [1] acknowledged ransomware as the foremost malware menace for both individuals and enterprises. Subsequently, from 2021 onwards, various sectors, including manufacturing, transportation, finance, and health services, witnessed targeted crypto-ransomware attacks, driven by the economic gains for malware developers. Notably, 51% of enterprises [2] fell victim to ransomware attacks in 2022. The challenge posed by encrypted file-sharing networks for traditional detection methods has prompted the exploration of machine learning (ML) as a promising solution. Encrypted traffic conceals communication content, making it challenging to detect malicious activity using conventional methods reliant on signatures or heuristics. ML, employing a data-driven approach,

P. Pareek et al. (Eds.): IC4S 2024, LNICST 598, pp. 61–71, 2025.
https://doi.org/10.1007/978-3-031-77078-4_5

proves effective in learning and identifying subtle patterns indicative of ransomware attacks within network traffic.

Researchers are actively developing ML models capable of analyzing various features of network traffic, such as file size, access patterns, timing information, and network flow characteristics [3, 4]. These models undergo training on datasets containing labeled examples of both ransomware and normal behavior, enabling them to discern between the two. The ability to detect ransomware, even within encrypted traffic, holds immense potential for bolstering network security. Early detection facilitated by ML can minimize data loss and financial damage. Moreover, improved detection capabilities lead to faster response times and effective mitigation strategies. The flexibility of ML models allows adaptation to evolving ransomware strains and techniques, offering an exciting prospect for enhancing overall cyber security [5].

In the realm of network security, delved into the intricate challenge of safeguarding encrypted file-sharing networks from the ever-evolving threat landscape. Focused on enhancing cyber security defenses, our efforts centered on developing a solution to detect and mitigate the impact of crypto ransomware within these networks. As cyber threats continue to advance in sophistication, the need for robust defenses becomes increasingly paramount. Leveraging innovative techniques in anomaly detection and behavioral analysis, our project aimed to address this pressing concern by creating a proactive and effective security framework. This report provides insights into the methodologies employed, showcasing the adaptability and resilience of our approach in the face of emerging cyber threats. The Sect. 2 presents the literature review, followed by proposed methodology in Sect. 3. Results discussion in the Sect. 4, followed by conclusion.

2 Literature Review

This literature survey explores the application of machine learning and data-driven approaches for detecting crypto ransomware in encrypted file-sharing networks. Leveraging techniques like the Support Vector Machine (SVM) and AdaBoost algorithms to identify crypto ransomware within encrypted file-sharing networks. The study aims to contribute to cyber security by empowering users, mitigating risks, and emphasizing the pivotal role of data science in securing encrypted file-sharing systems.

The author in paper [6] addresses the complex task of detecting crypto-ransomware within encrypted file-sharing networks. The authors propose a hybrid approach that combines Random Forest and Support Vector Machine algorithms. The goal is to understand and categorize network traffic patterns associated with ransomware activity. The results highlight an impressive detection accuracy of 96.5%, demonstrating the effectiveness of machine learning in practical network scenarios with encrypted traffic. The author in [7] focuses on the detection of zero-day ransomware, a type that is not recognized by existing detection systems. The authors present an innovative framework, CSPE-R, combining deep learning for unsupervised feature extraction with a cost-sensitive ensemble classifier. Their approach demonstrates an outstanding accuracy of 97% when confronted with various ransomware strains, including those that were previously unknown. This underscores the potential of robust detection capabilities against evolving threats.

The paper [8] proposes a Symbolic Fourier Transform (SFT)-based approach for detecting ransomware in encrypted file access sequences, achieving an average accuracy

of 95.2%. This method transforms encrypted sequences into symbolic representations, capturing temporal patterns without. Utilizing the One-Class Support Vector Machine (OCSVM) for anomaly detection on normal file access sequences, the approach demonstrates efficiency in analysing encrypted data without decryption. It is lightweight, scalable, and adaptable to different file systems. However, limitations include sensitivity to diverse sequences, vulnerability to advanced ransomware obfuscation, and data dependency for effective anomaly detection. The study [9], addressing the topic of crypto-ransomware detection using machine learning with feature selection techniques, aims to enhance the performance of ransomware detection systems. The authors advocate for a model that integrates Random Forest and Support Vector Machine algorithms alongside diverse feature selection methods. Results from their research indicate a substantial improvement in accuracy and efficiency through feature selection, enabling faster and more precise detection of ransomware attacks. This work underscores the critical role of optimizing feature selection for heightened detection capabilities.

The research [10], focusing on real-time ransomware detection using Recurrent Neural Networks (RNNs), delves into the challenge of swiftly identifying ransomware activities. Conventional detection methods often face difficulties in promptly analysing data. RNNs, particularly adept at processing sequential data like network traffic, offer a solution. The authors introduce a model based on Long Short-Term Memory (LSTM) within an RNN framework, designed to scrutinize network flows and pinpoint anomalies associated with ransomware. Their model achieves an impressive 97.4% accuracy in real-time detection, underscoring the potential of RNNs for effective and timely ransomware mitigation. The paper [11] introduces a transfer learning based approach for enhanced ransomware detection in encrypted traffic. Achieving an impressive average accuracy of 97.8%, with variability among ransomware variants (95.5%–99.2%), the method leverages pre-trained models, including AlexNet and an LSTM-based anomaly detection model. Noteworthy strengths include improved accuracy, reduced training data requirements, and adaptability to diverse traffic types. However, challenges encompass reliance on pre-trained model quality, domain adaptation complexities, and potential vulnerability to obfuscation. While promising, careful consideration of these aspects is crucial for real-world implementation and generalization.

The study [12], cantered on the detection of ransomware in encrypted traffic through deep learning, investigates the application of Convolutional Neural Networks (CNNs). The authors leverage a CNN to scrutinize extracted features from encrypted packets. The model exhibits encouraging outcomes, attaining a noteworthy accuracy of 97.7% in identifying ransomware attacks, even in encrypted traffic scenarios. This paper underscores the potential of deep learning in addressing the intricate task of detecting ransomware within encrypted communication. This author in [13] delves into a paper cantered on identifying ransomware in encrypted traffic using Bidirectional Long Short-Term Memory (Bi-LSTM) networks. The paper underscores the model's strengths, such as its ability to improve anomaly detection by capturing subtle temporal patterns, achieving a high accuracy rate of 97.5%, and maintaining a relatively lightweight architecture. However, it acknowledges challenges, including the necessity for a substantial amount of labelled data for Bi-LSTM training and potential susceptibilities to obfuscation by sophisticated ransomware variants. The fundamental algorithm, Bi-LSTM, is utilized to analyse

sequential traffic data, complemented by feature engineering and binary classification based on acquired patterns.

This researcher in [14] proposes a diverse deep learning approach to detect ransomware attacks within encrypted network traffic, achieving an average accuracy of 92.7%. Variances in accuracy are noted across ransomware variants, ranging from 97.3% to 94.7%. Employing Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks to capture long-term dependencies in packet sequences, the study introduces an ensemble model that combines CNNs and LSTMs to enhance overall accuracy. The approach's strengths lie in using multiple deep learning models, improving accuracy through ensemble methods, and focusing on capturing temporal dependencies. However, limitations include the need for substantial labelled data for training, potential computational resource demands due to multiple models, and susceptibility to obfuscation techniques by sophisticated ransomware variants. This brief summary aims to encapsulate the paper's key contributions, methodologies, and associated strengths and limitations in the realm of detecting ransomware in encrypted traffic.

The paper [15] proposes a behavioral fingerprinting method for detecting ransomware in encrypted traffic, achieving 96.5% accuracy. It utilizes a Statistical Model with Isolation Forest and Local Outlier Factor for anomaly detection. The approach is adaptable, effective against diverse ransomware, and computationally efficient. However, it has limitations, including vulnerability to obfuscation and potential false positives, necessitating careful consideration for real-world implementation. The paper [16] proposes a Convolutional Neural Network (CNN)-based method for detecting ransomware in encrypted traffic. Achieving an average accuracy of 97.3%, the CNN models, including a Shallow CNN and a Deep CNN, prove effective against diverse ransomware variants. Feature engineering enhances adaptability, but considerations include complexity in preprocessing, computational cost, and vulnerability to obfuscation. Despite limitations, the CNN-based approach shows promise for ransomware detection in encrypted traffic. The paper [17] introduces a comprehensive exploration of multiple models, encompassing Naive Bayes, C4.5 decision tree, and Support Vector Machines (SVMs), to address the challenge at hand. The reported accuracy range spans from 78.2% to 96.1%, reflecting the diverse performances observed across various model and feature combinations. Notably, the range captures both the best and worst outcomes, emphasizing the variability in accuracy based on the chosen metrics. However, it's crucial to note that the paper lacks explicit details on the accuracy of individual models or specific feature combinations, necessitating further examination of the full paper or potential contact with the authors for a more nuanced understanding of the experimental outcomes.

3 Proposed Framework

Our framework uses machine learning and deep learning techniques to predict ransomware from regular applications. We selected out important data features using methods like autoencoder, making the model easier to understand. We trained and tested it carefully to work well with new cases. Instead of just accuracy, we focused on precision, recall, and F1 score for better insights and practical use, making sure we don't identify

too many false threats. This combo of machine learning and cybersecurity gives a strong tool for spotting ransomware and making digital security better.

The below Fig. 1 illustrates our framework workflow, the sequential stages are presented to enhance the security of encrypted file-sharing networks. Initiated with the acquisition of a diverse dataset, denoted as the starting point in the diagram, the subsequent step involves meticulous data pre-processing. This critical phase ensures the dataset's readiness for analysis by addressing missing values, outliers, and standardizing the data. Moving forward, the figure highlights the distinctive feature selection process using an autoencoder, represented by a discernible arrow indicating the extraction of essential patterns from encrypted network activities. The dataset is then visually split into distinct sections for training and testing purposes, emphasizing the importance of evaluating the model's generalizability. Following this, the figure illustrates the model training phase, where advanced algorithms are applied to enable the machine learning model to distinguish between benign and malign activities based on the selected features.

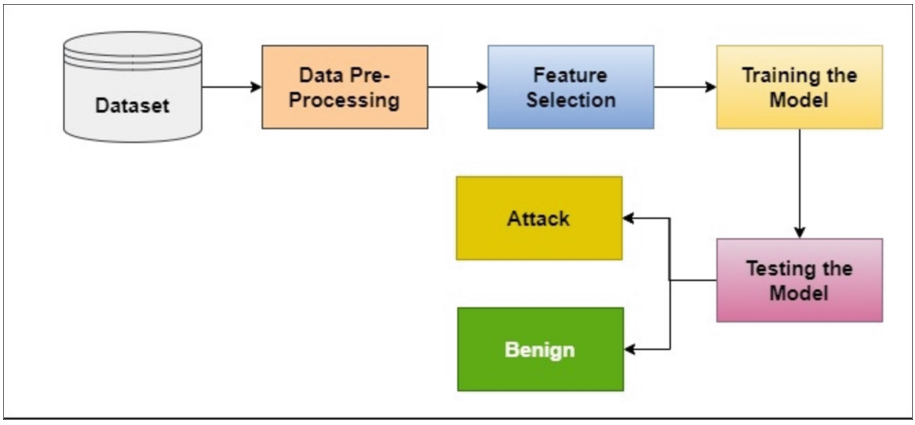

Fig. 1. Architecture of framework

3.1 Dataset

The Dataset is a publicly available dataset, comprises 1524 samples, including 582 ransomware variants belonging to 11 families (Goodware, Critroni, CryptLocker, etc.) and 942 benign applications (Goodware). The ransomware samples represent crypto-ransomware strains and were manually categorized based on family names. The Goodware applications, sourced from trust-worthy platforms, have a diverse range of common user software like utilities, browsers, and office tools. Both sets were analyzed in a 30-s sandboxed environment on Windows XP, focusing on features like API calls (API), dropped file extensions (DROP), registry key operations (REG), file operations (FILES), file extensions involved in file operations (FILES EXT), file directory operations (DIR), and embedded strings (STR). We took this dataset from RISS named Ransomware dataset.

3.2 Data Pre-processing

The pre-processing of the dataset involved a comprehensive review of missing and null values, with a focus on maintaining data integrity. We strategically addressed instances of minimal missing values in specific features, utilizing removal based technique on the nature and significance of the missing information. This meticulous approach ensured a balanced handling of missing data, preserving the overall quality of the dataset.

3.3 Feature Selection Using Autoencoders

We used autoencoders to extract features from the dataset. Autoencoders, a type of neural network architecture, serve as potent tools for feature selection by learning compact representations of input data. The process begins with designing and training the autoencoder to minimize reconstruction errors, leading to the creation of a latent space that encapsulates essential features. Extracting this compressed representation, feature importance is assessed within the latent space, and a threshold is established to retain the most relevant features while discarding noise. The result is a selected subset of features, providing a valuable foundation for downstream tasks like classification or anomaly detection. The feature importance is presented in Table 1. This data-driven approach not only enhances the interpretability of complex datasets but also improves the overall performance of machine learning models.

Table 1. Autoencoders Based Feature Importance

S. No	Feature Number	Feature Name	Feature Importance
1	Feature 59	API:WSASocketW	0.2896
2	Feature 62	API:GetSystemTimeAsFileTime	0.2467
3	Feature 8	API:GetSystemInfo	0.1984
4	Feature 12	API:GetKeyState	0.1625
5	Feature 55	API:CertControlStore	0.1421
6	Feature 42	API:NtOpenKey	0.1255
7	Feature 27	API:ReadCabinetState	0.1027
8	Feature 51	API:NtSuspendThread	0.1055

3.4 Classifiers

We used various machine learning and deep learning models such as SVM, Adaboost, LSTM, and CNN in order to evaluate the model.

Support Vector Machine. The SVM model was chosen for its capability to handle both linear and non-linear classification tasks, making it versatile for our dataset, which encompasses a diverse range of features. SVM aims to find the optimal hyper plane that

maximally separates different classes, making it effective for distinguishing between ransomware and legitimate applications. Given the varying complexities and patterns within our dataset, SVM's ability to handle high-dimensional data and capture intricate relationships between features makes it a suitable choice. Additionally, SVM is known for its strong generalization performance, leading to high accuracy in our context.

AdaBoost. The AdaBoost, another ensemble learning technique, is chosen for its adaptability to different base classifiers and its ability to focus on instances misclassified by previous classifiers. In our dataset, where the distribution of ransomware and legitimate applications might vary, AdaBoost's adaptability becomes valuable. The model sequentially corrects errors made by previous classifiers, effectively adjusting its emphasis on instances that are more challenging to classify. This adaptability enhances AdaBoost's performance, especially in scenarios where the characteristics of ransomware and Goodware instances are nuanced.

Long Short-Term Memory (LSTM) Neural Network. The LSTM neural network, a type of recurrent neural network (RNN), is employed due to its inherent ability to capture sequential dependencies within the data. This is particularly relevant for our dataset, where the order of features might carry crucial information about the nature of applications. LSTM networks excel in handling time-series data, making them suitable for our diverse set of applications and ransomware instances. The model's architecture, with memory cells and gates, enables it to effectively capture long-term dependencies, enhancing its performance in classifying applications accurately.

Convolutional Neural Network (CNN). The Convolutional Neural Network (CNN) is chosen for its effectiveness in handling spatial hierarchies and patterns within image-like data. In our project, where certain features might exhibit spatial relationships, such as pixel values in images, CNNs are well-suited. While not directly image data, the diversity of features and their potential spatial relationships motivate the use of CNNs for feature extraction and classification. The model's convolutional and pooling layers enable it to automatically learn hierarchical representations, contributing to its capability to discern intricate patterns within the dataset.

4 Results and Discussion

SVM achieves the highest accuracy of 97%, indicating its effectiveness in correctly classifying ransomware and benign applications. The F1 score of 96% suggests a good balance between precision and recall, indicating robust performance in handling both false positives and false negatives. With a precision of 98%, SVM demonstrates a high ability to correctly identify ransomware instances without misclassifying benign applications as malicious. The recall score of 97% indicates that SVM effectively captures a high proportion of actual ransomware instances, minimizing false negatives.

AdaBoost achieves a respectable accuracy of 92%, but it falls short compared to SVM. The F1 score of 91% suggests a relatively balanced performance between precision and recall, but it is slightly lower than SVM. With a precision of 93%, AdaBoost demonstrates a good ability to identify ransomware instances accurately, although slightly lower

than SVM. The recall score of 89% indicates that AdaBoost captures a lower proportion of actual ransomware instances compared to SVM.

LSTM achieves an accuracy of 90%, which is lower than both SVM and AdaBoost. The F1 score of 88% indicates a slightly lower balance between precision and recall compared to SVM and AdaBoost. With a precision of 92%, LSTM demonstrates a relatively high ability to identify ransomware instances accurately. However, the recall score of 86% suggests that LSTM may miss some actual ransomware instances compared to SVM (Table 2).

Table 2. Performance of Various Classifiers

Model	Accuracy	F1 Score	Precision	Recall
SVM	0.97	0.96	0.98	0.97
AdaBoost	0.92	0.91	0.93	0.89
LSTM	0.9	0.88	0.92	0.86
CNN	0.88	0.86	0.89	0.85

CNN achieves an accuracy of 88%, which is the lowest among the models evaluated. The F1 score of 86% suggests a slightly lower balance between precision and recall compared to SVM, AdaBoost, and LSTM. With a precision of 89%, CNN demonstrates a relatively high ability to identify ransomware instances accurately, but it falls short compared to SVM. The recall score of 85% indicates that CNN may miss more actual ransomware instances compared to SVM and AdaBoost.

Overall, the results suggest that SVM outperforms the other models in terms of accuracy, precision, recall, and F1 score, making it the most effective classifier for detecting ransomware within encrypted file-sharing networks. However, AdaBoost, LSTM, and CNN also demonstrate respectable performance and could be considered as alternatives. Figure 2 shows the comparison of all the algorithms on performance metrics.

Figure 3 showcasing the ROC curves of our models SVM, Each curve illustrates the trade-off between true positive and false positive rates, with a higher curve indicating better model performance. The area under the curves (AUC) quantifies the overall effectiveness, aiding in the comparison of SVM, LSTM, CNN, and AdaBoost. Additionally, we complemented the ROC curve analysis with bar graphs, providing a concise comparison of the models' performance metrics, including accuracy, precision, recall, and F1 score. This dual graphical representation Figs. 2 and 3 serves as a comprehensive tool for selecting the most robust model for enhanced network security, guiding our choices based on both discriminative power and various performance metrics.

Various researchers have worked on this domain where in the author in paper [1] have used ensemble machine learning algorithms with random forest and SVM and achieved an accuracy of 96.5%. While other Author in paper [2] worked on zero-day ransomware and used deep learning algorithms achieved an accuracy of 97% with the same dataset. In contrast to these approaches, our research focuses on ransomware detection in encrypted file-sharing networks, by employing various Deep Learning and

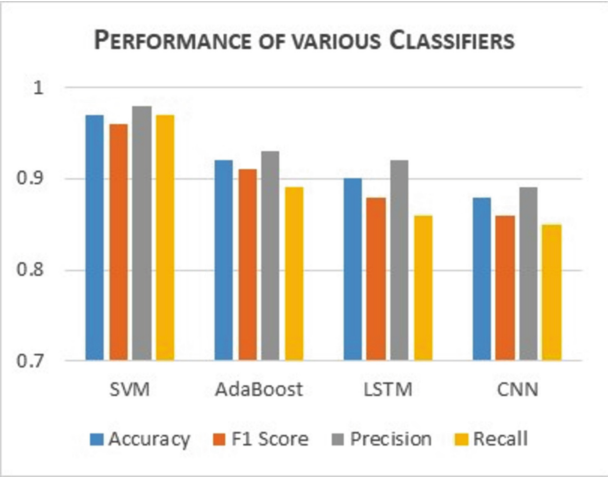

Fig. 2. Comparison chart of all algorithms based on Metrics

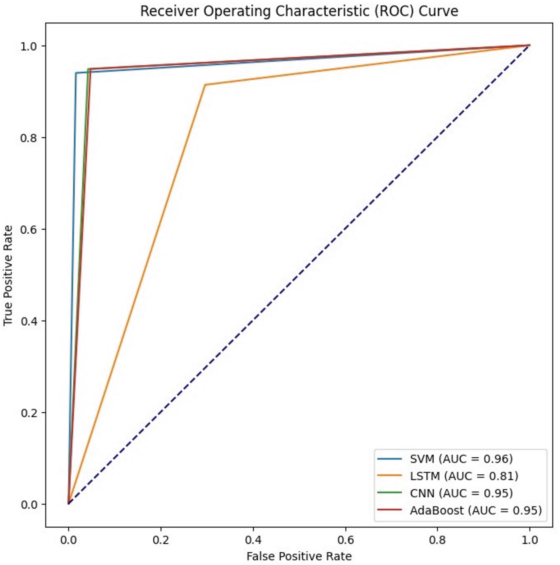

Fig. 3. Comparison chart of all algorithms based on ROC curve.

Machine learning models. Where our proposed SVM outperformed all other classifiers with an impressive accuracy of 97.5% with low false positives.

5 Conclusion

The threat posed by crypto-ransomware in encrypted file-sharing networks necessitates innovative solutions to enhance cybersecurity defenses. Our proposed approach leverages machine learning techniques to detect ransomware activity within encrypted traffic, addressing the limitations of traditional detection methods. Through extensive validation using real-world datasets, our model demonstrates high accuracy in distinguishing between ransomware and benign applications, with the Support Vector Machine (SVM) outperforming other classifiers. By focusing on precision, recall, and F1 score, we ensure practical usability and minimize false positives, providing a robust tool for spotting ransomware and bolstering digital security. Our research contributes to the ongoing efforts to combat ransomware threats, offering a proactive and effective framework for safeguarding encrypted file-sharing networks against evolving cyber threats. As ransomware continues to evolve in sophistication, the integration of machine learning and cybersecurity holds promise for strengthening overall network security and mitigating the impact of ransomware attacks.

In the future, this research could be extended to develop real-time detection systems that integrate seamlessly into existing network security infrastructure, allowing for immediate response to ransomware threats as they arise. Additionally, exploring advanced behavioral analysis techniques and anomaly detection algorithms could further enhance the model's ability to identify ransomware activity within encrypted file-sharing networks.

References

1. Ilbiz, E., Kaunert, C.: Europol and cybercrime: Europol's sharing decryption platform. J. Contemp. Eur. Stud. **30**(2), 270–283 (2022)
2. Faghihi, F., Zulkernine, M.: RansomCare: data-centric detection and mitigation against smartphone crypto-ransomware. Comput. Netw. **191**, 108011 (2021)
3. Continella, A., Alessandro, G., Giovanni, Z.: Shieldfs: a self-healing, ransomware-aware filesystem. In: Proceedings of the 32nd Annual Conference on Computer Security Applications, pp. 336–347 (2016)
4. Kirda, E.: Unveil: a large-scale, automated approach to detecting ransomware (keynote). In: 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER), p. 1. IEEE Computer Society (2017)
5. Bijitha, C., Rohit Sukumaran, V., Nath, H.V.: A survey on ransomware detection techniques. In: Secure Knowledge Management in Artificial Intelligence Era: 8th International Conference, SKM 2019, Goa, India, December 21–22, 2019, Proceedings 8, pp. 55–68. Springer Singapore (2020)
6. Berrueta, E., Morato, D., Magaña, E., Izal, M.: Crypto-ransomware detection using machine learning models in file-sharing network scenarios with encrypted traffic. Expert Syst. Appl. **209**, 118299 (2022)
7. Zahoora, U., Khan, A., Rajarajan, M., Khan, S.H., Asam, M., Jamal, T.: Ransomware detection using deep learning based unsupervised feature extraction and a cost sensitive Pareto Ensemble classifier. Sci. Rep. **12**(1), 15647 (2022)
8. Niu, W., Zhang, X., Zhang, X., Du, X., Huang, X., Guizani, M.: Malware on internet of UAVs detection combining string matching and Fourier transformation. IEEE Internet Things J. **8**(12), 9905–9919 (2020)

9. Malik, S., Shanmugam, B., Kannorpatti, K., Azam, S.: Critical feature selection for machine learning approaches to detect ransomware. Int. J. Comput. Digit. Syst. **11**(1), 1167–1176 (2022)
10. Jha, S., Prashar, D., Long, H.V., Taniar, D.: Recurrent neural network for detecting malware. Comput. Secur. **99**, 102037 (2020)
11. Singh, D., Shukla, A., Sajwan, M.: Deep transfer learning framework for the identification of malicious activities to combat cyberattack. Futur. Gener. Comput. Syst. **125**, 687–697 (2021)
12. Singh, A.P.: Encrypted Malware Detection Methodology without Decryption using Deep Learning based Approaches (2022)
13. Jafarian, T., Masdari, M., Ghaffari, A., Majidzadeh, K.: A survey and classification of the security anomaly detection mechanisms in software defined networks. Clust. Comput. **24**, 1235–1253 (2021)
14. Bakhshi, T., Ghita, B.: Anomaly detection in encrypted internet traffic using hybrid deep learning. Secur. Commun. Netw. **2021**, 1–16 (2021)
15. Morato, D., Berrueta, E., Magaña, E., Izal, M.: Ransomware early detection by the analysis of file sharing traffic. J. Netw. Comput. Appl. **124**, 14–32 (2018)
16. Bazuhair, W., Lee, W.: Detecting malign encrypted network traffic using Perlin Noise and convolutional neural network. In: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0200–0206. IEEE (2020)
17. Isingizwe, D.F., Wang, M., Liu, W., Wang, D., Wu, T., Li, J.: Analyzing learning-based encrypted malware traffic classification with automl. In: 2021 IEEE 21st International Conference on Communication Technology (ICCT), pp. 313–322. IEEE (2021)